BRILL

CAHIERS DE LINGUISTIQUE ASIE ORIENTALE
EAST ASIAN LANGUAGES AND LINGUISTICS
52 (2023) 137–162

CAHIERS DE LINGUISTIQUE ASIE ORIENTALE
brill.com/clao

Evaluating rhyme annotations for large corpora

Metrics and data

Julien BALEY | ORCID: 0000-0003-1056-6211
SOAS University of London, London, UK
julien.baley@gmail.com

Received 28 June 2022 | Accepted 30 May 2023 |
Published online 27 November 2023

Abstract

Recent methods have been proposed to produce automatic rhyme annotators for large rhymed corpora. These methods, such as Baley (2022b) greatly reduce the cost of annotating rhymed material, allowing historical linguists to focus on the analysis of the rhyme patterns. However, evidence for the quality of those annotations has been anecdotal, consisting of a handful of individual poem case studies. This paper proposes to address the issue: first, we discuss previously proposed metrics that evaluate the quality of an annotator's output against a ground-truth annotation (List, Hill, and Foster; 2019) and we propose an alternative metric that is better suited to the task. Then, sampling from Baley's published annotated corpus and re-annotating it by hand, we use the sample to demonstrate the lacunae in the original approach and show how to fix them. Finally, the hand-annotated sample and source code are published as additional data, so that other researchers can compare the performance of their own annotators.

Keywords

data annotation – evaluation metric – Chinese rhymes – Middle Chinese phonology

Résumé

Des méthodes ont récemment été proposées afin de développer des annotateurs automatiques de rime pour de larges corpus rimés. Ces méthodes, telles que celle présentée

Published with license by Koninklijke Brill NV | DOI:10.1163/19606028-bja10032

dans Baley (2022b), permettent de grandement réduire le coût d'annotation des textes rimés, permettant aux linguistes historiques de se concentrer sur l'analyse des motifs rimés résultants. Cependant, les preuves de la qualité de ces annotations sont anecdotiques, consistant en une poignée d'études de cas de poèmes. Cet article propose d'aborder ce problème: tout d'abord, nous discutons des métriques proposées précédemment qui évaluent la qualité des annotations produites par un annotateur par rapport à une annotation considérée comme exacte (List, Hill, and Foster (2019)) et nous proposons une métrique alternative qui, à notre avis, est mieux adaptée à la tâche. Ensuite, nous échantillonnons à partir du corpus annoté publié par Baley et le réannotons à la main. Nous utilisons l'échantillon pour démontrer les lacunes de l'approche d'origine et montrer comment les corriger. Enfin, l'échantillon annoté à la main est publié en tant que données supplémentaires, afin que d'autres chercheurs puissent comparer les performances de leurs propres annotateurs.

Mots-clés

annotation de données – métrique d'évaluation – rimes du chinois – phonologie du chinois moyen

1 Introduction

The study of Chinese rhymed material has long been of interest to historical linguistics trying to reconstruct the phonological system of the Chinese language: since the script does not explicitly indicate pronunciation, rhyming texts such as poems allow historical linguists to infer phonetic similarities between characters based on their frequent rhyming in texts.

To facilitate the analysis of such texts, List, Hill, and Foster (2019) have established an annotation standard that would allow the research community to develop a shared corpus of annotated poetry; List (2019) has also developed tools to help speed up hand-annotation efforts.

The extant Chinese rhyme corpus contains hundreds of thousands of texts, both poetry and rhymed prose, ranging from Shāng 商 dynasty (16th century BC–1046 BC) bronze vessels, all the way to the present. From the Hàn 漢 dynasty (202 BC–220 AD) onwards, the extant corpus of each period contains thousands of pieces. Until recently, the size of such a corpus made it too expensive to annotate and analyze. Re-using List's (2016) idea of using graph community detection algorithms to analyze rhyme patterns, Baley (2022b) introduced an approach to automatically discover rhyming character sets and automati-

cally annotate large rhymed corpora. This approach allows us to produce standard annotations of the various corpora mentioned above and focus on their analysis.

One issue with Baley (2022b) is that it is not known how correct the annotations are: although the approach is shown to work in a few case studies (and—in one instance—to fail, [Baley, 2022b: 67]), it offers no evidence for its accuracy across a large corpus. The present article has two goals: first, we want to evaluate whether the automatic annotation approach presented above can be relied on, and second, we want to offer the possibility for competing approaches to be compared with each other, so that we can arrive at higher-quality, standardized annotations of the entire extant Chinese rhymed corpus. To this aim, the article is structured around two main sections: the first section discusses how annotation quality should be measured, evaluate existing metrics proposals, and propose an alternative metric. In the second section, a sample from the annotated corpus published in Baley (2022a) is re-annotated by hand. The sample can serve as a standardized test set to compare competing automatic annotation approaches, and it is used to demonstrate lacunae in the earlier approach; I then propose a way to address those issues and demonstrate their efficacy. The final section offers suggestions on how to sample from other major corpora, so as to build a set of statistically useful corpora for future annotator evaluations.

2 Annotation accuracy metric

Using the annotation standard developed by List, Hill, and Foster (2019), we need a metric that compares two different annotations of a poem. If one of these annotations is considered a reliable "ground truth", then it can be used to assess the quality of the other annotation. The principle can be extended to a corpus of poems that we annotate by hand and use as a reference to evaluate various automatic annotation strategies. The metric is introduced in two parts: the first part considers how the problem of scoring rhyme judgement is similar to that of scoring a graph clustering and reviews the proposal made by List, Hill, and Foster (2019:40) to use B-cubed metrics. Then, the second part proposes an alternative scoring strategy and demonstrates how it addresses the limitations of the B-cubed metrics for our scenario.

2.1 *Rhyme judgement as a clustering problem and metric*

List, Hill, and Foster (2019:40) argue that "the assessment for a given stanza, whether two words rhyme or not, can also be thought of as a clustering task".

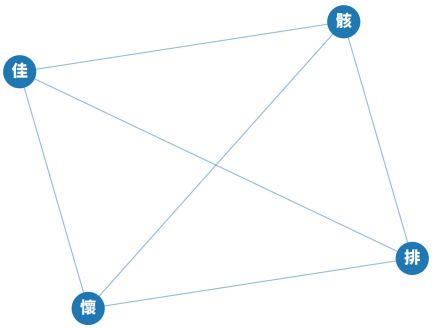

FIGURE 1
Graph of rhymes in Zhāng Dàoqià's poem, according to this article's author

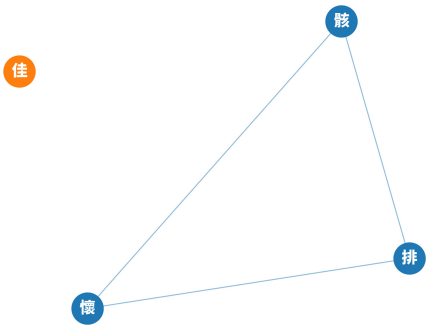

FIGURE 2
Graph of rhymes in Zhāng Dàoqià's poem, according to the Community annotator

TABLE 1 Méihuā èrshí shǒu: qí yībā 梅花二十首其一八 (Twenty Poems on Plum Blossoms, part 18) by Zhāng Dàoqià 張道洽 (1205–1268) (QSS 3293.39249)

Poem	Rhyme	LMC	Ground truth annotation	Baley's Community annotator
幾年冷樹雪封骨，一夜東風春透懷。	懷	xɦwa:j	a	a
花裏清含仙韻度，人中癯似我形骸。	骸	xɦja:j	a	a
三點兩點淡尤好，十枝五枝疏更佳。	佳	kja:j	a	b
野意終多官意少，玉堂茅舍任安排。	排	pɦa:j	a	a

The intuition behind this idea is best demonstrated using graphs. In Table 1, an example poem is taken from Baley (2022b: 75) for which the Community annotator fails, along with a "ground truth" annotation for that poem.

These two annotations correspond to two different partitionings of the {懷, 骸, 佳, 排} rhyme character set, as illustrated in Figure 1 and Figure 2.

Since rhyme judgement is similar to a clustering task, scoring the quality of a rhyme judgement against a ground truth is similar to scoring agreement between two partitions of a given node set. This is an extensively studied problem with many proposed metrics. List, Hill, and Foster (2019) propose to use B-cubed metrics, following Amigó et al. (2009) who demonstrate that it is the only metric that fulfills a set of constraints that they deem useful. The idea behind B-cubed metrics, from Bagga and Baldwin (1998), is to compare for each node of the graph how it has been clustered in the evaluated partitioning vs. in the ground truth partitioning, and treat it as an information retrieval task:

does the evaluated cluster contain all the elements of the ground truth cluster (recall) and vice versa (precision).

In mathematical terms, the definitions of B-cubed precision and recall are articulated around the concept of correctness between two elements *e* and *e'*: the relation between *e* and *e'* is considered correct if their sharing a category is correlated with their sharing a cluster:

$$Correctness(e, e') = \begin{cases} 1 \; iff category(e) = category(e') \leftrightarrow cluster(e) = cluster(e') \\ 0 \; otherwise \end{cases}$$

The B-cubed precision of a single element *e* is then defined as the average correctness between *e* and all the elements of its cluster, while recall considers the average correctness between *e* and all the elements of its category:

$$Precision(e) = Avg_{e', cluster(e)=cluster(e')}[Correctness(e, e')]$$

$$Recall(e) = Avg_{e', category(e)=category(e')}[Correctness(e, e')]$$

And the overall precision and recall of the clustering are defined as the average precision and recall over all the elements:

$$Precision = Avg_e[Precision(e)]$$

$$Recall = Avg_e[Recall(e)]$$

In the case of the *hwɛj* 懷 node, the Community annotator clusters it with *hɛj* 骸 and *bɛj* 排 but not with *kɛ* 佳, while the ground truth annotation clusters them all together. This gives *hwɛj* 懷 a recall of 0.75 (of the 4 nodes found in the ground truth cluster containing *hwɛj* 懷, 3 are found in the Community cluster containing *hwɛj* 懷) and a precision of 1.0 (of the 3 nodes found in the Community cluster containing *hwɛj* 懷, all are found in the ground truth cluster containing *hwɛj* 懷). *hɛj* 骸 and *bɛj* 排 obtain identical scores by symmetry and *kɛ* 佳 gets a B-cubed recall of 0.25 and a precision of 1.0. A global score is produced by averaging over the four characters, which gives a B-cubed recall of $\frac{3\times0.75+0.25}{4} = 0.625$, a B-cubed precision of 1.0, and a single metric is obtained by computing the harmonic mean of recall and precision $F_1 = 2\times\frac{precision\times recall}{precision+recall} = 2\times\frac{0.625\times1.0}{0.625+1.0} = 0.769$.

TABLE 2 Second stanza of Juǎn'ěr 卷耳 (Rough Cocklebur) as annotated by Baxter (1992:584) and Wáng Lì (2014:138–139) (reconstructions are quoted from the original works)

Baxter annotation	Reconstruction	Wáng Lì annotation	Reconstruction
陟彼[a]崔[a]嵬	*Sduj, *nguj	陟彼崔[a]嵬	*nguəi
我馬[a]虺[a]隤	*xuj, *luj	我馬虺[a]隤	*duəi
我姑酌彼金[a]罍	*C-ruj	我姑酌彼金[a]罍	*luəi
維以不永[a]懷	*gruj	維以不永[a]懷	*hoəi

2.2 *The limits of the clustering analogy*

On the surface, rhyme judgements do behave like a clustering problem, as illustrated in the previous examples. In many poems, this is a fine analogy, but it relies on the assumption that the two rhyme judgements (the one under evaluation and the ground truth) consider the same set of characters; in our previous example, both the ground truth annotation and the Community annotator produce a partitioning graph involving the four characters {懷, 骸, 佳, 排}. This assumption considers that annotators agree *a priori* on what can rhyme, an assumption which does not hold in the general case. For instance, the community-annotated corpus published in Baley (2022a) only contains annotations for even-numbered lines, ignoring odd-numbered lines that frequently contribute to the rhyme scheme[1] and that a human annotator would take into account; for such poems, the two annotators produce a partition of non-identical sets of characters.

Closer to the original proposal by List, Hill, and Foster (2019) to use B-cubed metrics, the rhyme judgements of William Baxter and Wáng Lì on the rhymes of the Shījīng do sometimes differ regarding which characters are rhyming, such as in the second stanza of the third poem of the *Airs of the States* (*Guó fēng* 國風), *Juǎn'ěr* 卷耳 (Rough Cocklebur), shown in Table 2. While Baxter and Wáng agree in annotating the last character of all four lines as rhyming, Baxter additionally annotates the penultimate characters of the first two lines as rhyming with the four others. Following the graph clustering analogy, this corresponds to Figure 3 and Figure 4.

This is where the analogy breaks: to evaluate the similarity between two clusterings of a set of nodes, the two clusterings must partition the same set of

1 Typically, in regulated verse, in addition to the last characters of lines 2 and 4, quatrains often have the last character of line 1 rhyming.

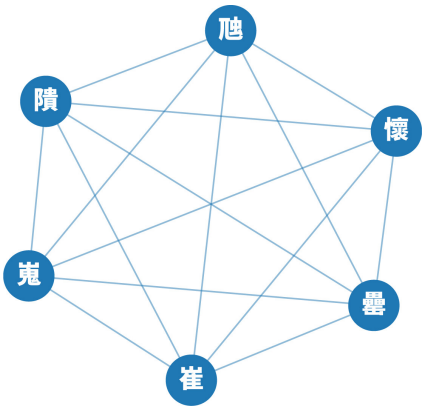

FIGURE 3
Juǎn'ěr 卷耳 as annotated by Baxter

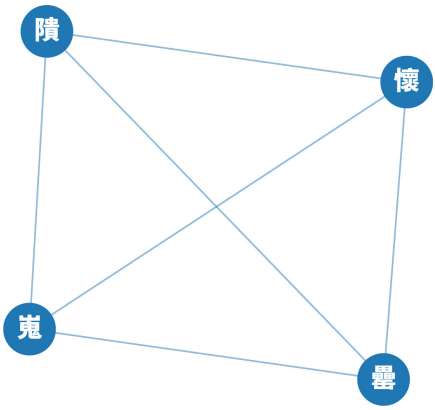

FIGURE 4
Juǎn'ěr 卷耳 as annotated by Wáng Lì

nodes; when two annotations differ in their choice of which characters rhyme, the metric cannot be computed. The simplest way to address this issue is to supplement annotations by taking the superset of characters annotated in either annotation, and if an annotation of a poem is missing an annotation for a given character, we add a new mark to that character. For Wáng Lì's annotation of the *Juǎn'ěr*, this means we need to annotate 崔 and 虺, respectively as [b] and [c] to indicate that they do not rhyme (in Wáng's annotation) with the other four characters. This gives Table 3 and Figure 5.

With this simple strategy, it is now possible to compute the B-cubed metrics of Wáng against Baxter for this stanza: B-cubed recall = 0.5, B-cubed precision = 1.0, B-cubed F_1 = 0.667 (if measuring Baxter against Wáng, we simply swap recall and precision; F_1 is not affected).[2]

2.3 *The issue with B-cubed metrics*

Whereas the issue above is unrelated to the choice of metrics per se and is easily solved, the issue presented below is inherent to B-cubed metrics and can only be addressed by using a different metric.

For the sake of the argument, we imagine a third human annotator who would take Baxter's annotation as a basis and notice that the first character

2 The code published by List et al. (in their original paper, and as of 2023/05/01) to produce B-cubed scores returns a score of 1.0 for this stanza, suggesting an error in the code (or that it only considers the last character of a line). We suggest an alternative implementation should be used until this is addressed.

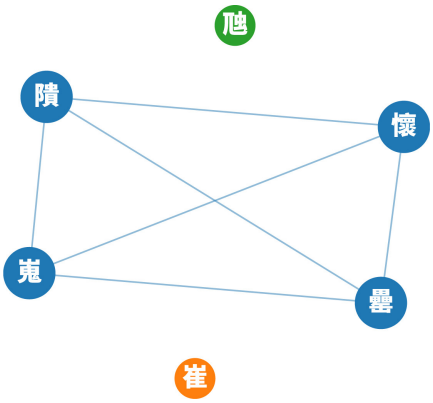

FIGURE 5
Aligned annotation of Juǎn'ěr 卷耳 by Wáng Lì

TABLE 3 Aligned annotations of Juǎn'ěr 卷耳 by Baxter and Wáng Lì

Baxter	Wáng Lì (aligned)
陟彼[a]崔[a]嵬	陟彼[b]崔[a]嵬
我馬[a]虺[a]隤	我馬[c]虺[a]隤
我姑酌彼金[a]罍	我姑酌彼金[a]罍
維以不永[a]懷	維以不永[a]懷

of the last line, 維, might also rhyme.[3] Since neither Baxter nor Wáng annotate this character, in order to score this third annotator against any of them the alignment technique described above needs to be applied, producing Table 4, and the B-cubed metrics can then be computed, as in Table 5.

The performance of this third annotator can be reported as 0.56 or 0.86 depending on whether one regards respectively Wáng or Baxter to be correct. The surprise, here, is that the Wáng / Baxter score which was reported as 0.667 in the previous section is now 0.73. The score increased because when comparing Wáng and Baxter on the 7-character set (instead of the 6-character set previously), Baxter and Wáng agree on 維 being in a cluster of its own (i.e. not rhyming with anything else) and this agreement is reflected in the B-cubed recall increasing (precision is already 1.0) and F_1 too.

3 On the basis of Jacques (2000) who demonstrates that *ywij* 維 originally had an **-uj* rhyme, Baxter and Sagart (2014: 366) reconstruct *ɢʷ*uj* as a stage earlier than Old Chinese *ɢʷij. This means *ɢʷ*uj* 維 could have rhymed with the other *-*uj* characters of the poem.

TABLE 4 Aligned annotations of Juǎn'ěr 卷耳 by Baxter, Wáng Lì and a third annotator

Baxter (aligned)	Wáng Lì (aligned)	Third annotator
陟彼[a]崔[a]嵬	陟彼[b]崔[a]嵬	陟彼[a]崔[a]嵬
我馬[a]虺[a]隤	我馬[c]虺[a]隤	我馬[a]虺[a]隤
我姑酌彼金[a]罍	我姑酌彼金[a]罍	我姑酌彼金[a]罍
[b]維以不永[a]懷	[d]維以不永[a]懷	[a]維以不永[a]懷

TABLE 5 B-cubed F_1 score between Baxter, Wáng Lì and the third annotator

Annotator pair	B-cubed F_1 score
Third annotator / Baxter	0.86
Third annotator / Wáng	0.56
Wáng / Baxter	0.73

This is a serious problem, because it means that any report of a B-cubed score is meaningless unless one fully specifies which set of characters were annotated, for the entire corpus. Additionally, this means that scores reported in different publications cannot be compared; instead, any new rhyme annotation of a corpus must gain access to all previously published annotations of the same corpus, align them jointly (and not by pairs) and produce a new set of results that supersedes the previous publications.

The only way around this issue is to require all reports of B-cubed metrics to be based on fully annotated poems, as demonstrated for Baxter's annotation in Table 6. As shown above, since adding padding annotations mechanically increases the B-cubed metrics, such maximally annotated poems would tend to produce a very high score, and although these would now be comparable (and would not evolve across time), one would not gain much intuition from their values as the difference between very poor annotators and perfect one would be small and mainly dependent on the size of the poem (heptasyllabic poems scoring higher than pentasyllabic ones), and the difference between several more realistic annotators would be minute.[4]

4 Under such a scheme, the B-cubed F_1 between Wáng and Baxter for this stanza jumps to 0.91;

TABLE 6 Maximally annotated extension of Baxter's annotation of Juǎn'ěr 卷耳

Baxter (maximally aligned)
[b]陟[c]彼[a]崔[a]嵬
[d]我[e]馬[a]虺[a]隤
[f]我[g]姑[h]酌[i]彼[j]金[a]罍
[k]維[l]以[m]不[n]永[a]懷

2.4 *Proposal for a rhyme annotation metric*

The set of constraints required to be able to report B-cubed metrics in the context of rhyme annotations as well as the difficult interpretation of those results make B-cubed metrics impractical and undesirable for measuring annotation quality. The following section proposes a metric that addresses these issues and is used in the rest of the article.

First, note that the problem with B-cubed metrics is shared by all metrics which focus on individual nodes of the graph (and produce a score per node, which is then averaged across the corpus): any such metric will suffer from the alignment problems presented above,[5] and we must therefore look elsewhere. If a node-based metric cannot be used, then the obvious alternative is to use an edge-based metric, i.e. the links between two nodes that indicate that two characters are rhyming. Within those parameters, the specific metric can be a matter of choice; here, I propose a very simple one and prove that it has the desired properties. I invite future research to propose better alternatives.

The idea behind the proposed metric is to take the set of edges from the rhyme annotation graph, and compute the traditional precision, recall and F_1 scores. Referring to our previous example, the Baxter graph in Figure 3 contains 15 edges, one between each pair of characters, while Wáng's graph in Figure 5 contains 6 edges. Recall is then calculated as representing the ratio of shared edges by the size of the reference edge set, and precision is the ratio of shared edges by the size of the evaluated edge set. As with B-cubed metrics, swap-

this can be compared to a dummy annotator that considers that no character rhymes with any other: its score against Baxter would be 0.85.

5 Identifying that B-cubed metrics tend to produce artificially high scores, Van Heusden et al. (2022) propose to only include clusters of more-than-1 elements in the computation. This does indeed produce lower scores, but the problem of comparability of the results remains.

ping which annotation is the reference and which is evaluated swaps precision and recall, and F_1—being the harmonic mean of precision and recall—remains unchanged. For Wáng evaluated against Baxter, since Wáng's edge set is a strict subset of Baxter's edge set, this gives a precision of 1.0 (all rhymes identified by Wáng are found in Baxter's annotation) and a recall of 0.4 (only 6 out of 15 rhyme pairs identified by Baxter were also identified by Wáng), with F_1 being 0.57.

This metric does not depend on a particular alignment: whether one chooses Figure 4 or Figure 5 as Wáng's annotation, although their node set is different, their edge set is identical and therefore so is their score. This means that this metric can be reported on its own, without any further qualification, and can be quoted in later publications as is.

This metric belongs to the family of pair counting metrics. In their comparison of clustering metrics, Amigó et al. (2009:11) prove that pair counting metrics only satisfy two of their four desirable constraints, namely:

– it satisfies the "cluster homogeneity" constraint: it is preferable for a cluster of the annotation under evaluation to only contain elements that do rhyme according to the ground truth. (this is related to precision)
– it satisfies the "cluster completeness" constraint: it is better to group elements that do rhyme according to the ground truth in a single cluster. (this is related to recall)

The other two constraints are not satisfied:

– it does not satisfy the "rag bag" constraint that states that if a character rhymes with nothing else in the poem, it is preferable to misclassify it as rhyming with other characters that don't rhyme with anything than to misclassify it as rhyming with a large group of inter-rhyming characters. Like all pair counting metrics, our metric simply has no such preference: misclassifying in one direction (towards a very clean cluster) or the other (towards an already noisy cluster) makes no difference to the score. In my opinion, this constraint—which was designed for information retrieval scenarios (search engine results)—does not add value in the context of scoring rhyme annotations. Therefore, whether a metric satisfies this constraint or not is irrelevant.
– it does not satisfy the "cluster size vs. quantity" constraints, which states that it is preferable to make a single mistake in a large cluster than many mistakes in small clusters. The metric does not satisfy this constraint due to the number of edges in a cluster being a quadratic function of the number of nodes in that same cluster.[6] B-cubed metrics, being node-based, do not suffer from

6 $N_{edges} = O(\frac{N_{nodes}(N_{nodes}-1)}{2})$

this quadratic bias for larger clusters and therefore satisfy this constraint. As opposed to the "rag bag" constraint, I do think this constraint is relevant to rhyme annotation scoring and it would be desirable to satisfy it; however, this constraint is at odds with the constraint of a metric being independent of the alignment, and I consider the latter more important to satisfy.

It is also worth noting that the "cluster size vs. quantity" constraint is particularly a problem when faced with sets of very unbalanced sizes, where some clusters are much larger than others. In the dataset from Baley (2022a), the overwhelming majority of poems have a single rhyme and therefore no imbalance; for the minority of poems that show several rhymes, the average ratio of "biggest cluster size" to "smallest cluster size" is 2.2, which is not a very large imbalance. This means that, in practice, not satisfying this constraint is not overly problematic for the dataset under consideration, but this could vary by dataset. When aggregated over a corpus, however, this metric gives more weight to longer poems; this seems acceptable and perhaps even desirable, as intuitively annotating very short poems is trivial[7] and annotating long poems correctly is harder. This also means that mistakes in large poems are very heavily penalized, which also seems desirable: since annotators are already good, a challenging metric is preferable.

Overall, the advantages of this metric are far more desirable than its flaws are detrimental and it is used in the rest of this article. This is however a topic that would benefit from further research.

3 Hand-annotated sample corpus

In Baley (2022b), I claimed that this automated approach could significantly speed up annotation efforts, but only offered anecdotal evidence regarding the quality of the output. To evaluate the output of an automatic rhyme annotator, a hand-annotated, reliable annotation of the same material is needed. In the present case, since the corpus published in Baley (2022b) contains around 250,000 poems,[8] it is only practical to manually annotate a sample. The sample needs to be large enough to allow comparisons of competing annotators (e.g. difference in F_1 scores) to carry statistical significance. Based on my experience, a sample size of 400 poems is more than enough and allows for a diverse sample. The source code for the sampling and evaluation of the annotator is made

7 If a poem only has 2 lines, they must practically always rhyme.

8 These represent the *shī* poetry of the Táng 唐 (618–907) and the Sòng 宋 (960–1279) as collected in the Quán Táng Shī 全唐诗 (1979) and Quán Sòng Shī 全宋詩 (1998).

available in Baley (2022d) for reproduction of the results as well as application to other datasets.

3.1 *Sampling*

In Baley (2022b), three annotators are compared: one that learns rhyming patterns through community detection, one based on the *Guǎngyùn* 廣韻 and a 'naïve' one that assumes the last characters of every even-numbered line always rhyme together. Based on these three annotations, 5 combinations are analyzed: when the three annotators produce the same output; when they all produce a different output; and 3 cases where one of the three annotators differ from the other two. As this segmentation was at the heart of the case studies presented and the argumentation, when sampling from each of these categories, a minimum number of poems is guaranteed: first, the number of poems to be sampled from each category is based on their prevalence in the corpus (e.g. "the three annotators agree" covers 63.2% of poems, therefore we sample $0.632 \times 400 = 253$ poems from this category); then, to guarantee a minimum of poems is sampled in each category, at least 30 poems are sampled even if the prevalence-based number is lower.[9]

This approach creates an imbalance in the sample, since categories such as "Community disagrees" are oversampled. There are two ways to resolve this imbalance: the first one, adjusting the sampling size of the other categories so that the size of each category in the sample is in proportion to its size in the full corpus, would be intractable and would defeat the purpose of sampling.[10] The alternative approach is to keep these sampling sizes, but to adjust the computation of the scores to take oversampling into account: the scores are computed 'by category' and then weighted according to their prevalence in the full corpus, so that contributions of "Community disagrees", for instance, would correctly represent 0.14% of the total score. This preserves the small size of the sample and is the approach taken here, giving the number of poems found in Table 7, yielding 444 poems in total. For each category, the desired number of poems is sampled at random.

9 An arbitrary, conventional value for minimal sampling size, that is practical for us in this case.

10 Since, on a prevalence basis, "Community disagrees" would only need 0.6 poems to be sampled, using a minimum of 30 gives a $\frac{30}{0.6} = 50$ oversampling ratio. Using this ratio, 12,650 "All agree" poems ($= 253 \times 50$) would need to be sampled, and a total of 20,080 poems across all categories.

TABLE 7 Sample size by annotator agreement category

Category	Total poems	Ratio in corpus (%)	Sample size (w/ prevalence)	To annotate (with minimum 30)
All agree	152893	63.2	253	253
Naïve disagrees	12437	5.1	21	30
Guǎngyùn disagrees	60734	25.1	101	101
Community disagrees	334	0.14	0.6	30
All disagree	15830	6.5	26	30

3.2 *Hand annotation*

The sampled poems have then been manually annotated following the annotation standard introduced in List, Hill, and Foster (2019). Instead of starting from bare poems and annotating from scratch, the output of the Community annotator has been assumed to be generally correct enough and has been used as a starting point.[11]

The process of annotation by hand then consists of reviewing and amending the annotations provided by the automatic annotator, in effect resulting in a semi-automatic process. The more correct the original automatic annotation is, the faster the review process. This is interesting to researchers as it means that, even assuming that automatic annotators can never be perfect and that corpora need to be annotated by hand, the hand annotation process can be made significantly cheaper. In the present case, it took a total of 2h24' to annotate the 4004 lines contained in the 444 poems, i.e. an average of 27.8 lines per minute or 3.1 poems per minute.[12] Unsurprisingly, most of this time was spent

11 Since the previously published corpus did not include annotation of odd-numbered lines, which are often rhyming in *shī* poetry (usually the lines 1, 2 and 4 of a quatrain can rhyme; here, all lines are considered), an annotation of these lines has been automatically inferred based on the rest of the corpus: if two characters are annotated as rhyming in another poem, they are automatically annotated as rhyming in the current poem. This is a reasonable guess (it is generally correct, but might not always be) that allows to quickly fill this lacuna of the original publication and save time in the subsequent hand annotation process: it is faster to fix incorrect annotations than add missing ones.

12 An anonymous reviewer of this paper suggested that such an approach might be biased towards the original output of the Community annotator and not be correct, and further wondered what an inter-(human)-annotator agreement would look like. To test this, a sub-sample of 44 poems was hand-picked (out of the 444 annotated through the semi-automatic process), all annotations were removed and a colleague annotated the poems. The poems were hand-picked with the aim of being challenging for annotators, choos-

annotating long poems with more complex rhyme structures that had been poorly annotated by the Community annotator, and easier poems were significantly faster to annotate,[13] which is how it should be: the automation affords the researcher to spend time resolving the more difficult problems.

The hand-annotated corpus is made available to the community in Baley (2022c) so that competing automatic annotators can be produced. It would be useful for researchers to produce similarly hand-annotated corpora for other periods and genres of rhyming material.

4 Evaluation results

Using the hand-annotated corpus and the metric presented above, the annotators published in Baley (2022b)—the so-called "Naïve", "Guǎngyùn", and "Community" annotators—are evaluated. All the poems of the automatically annotated corpus are collected and the ones that appear in the hand-annotated corpus are retained. Then, for a given automatic annotator and a given poem, the annotations produced by the automatic annotator with those found in the manual annotation are aligned. The algorithm then produces the list of all possible edges between the annotated nodes and produce binary lists of rhyme judgements for all poems—1 indicating the presence of an edge between two characters and 0 its absence—so as to form one long list on which the recall, precision and F_1 score are computed.

Since the automatic annotations previously published only cover the even-numbered lines but the present hand-annotation also identifies rhymes in odd-numbered lines and possibly inside of lines, we examine three sets of results:

– The scores produced by strictly evaluating what was published, against the hand annotation; the results will reflect the lack of annotation of odd-numbered lines.
– The scores produced by only considering even-numbered lines (and discard the odd-numbered lines' annotations from the hand-annotated corpus); this

ing mainly poems that had very unusual patterns. The sub-sample as annotated by the present author and the colleague can be found in the released dataset. Using the metric, the inter-annotator F_1 score is 0.989. This can be contrasted with the much lower scores obtained by the automatic annotators against our own annotation for this sample: Naïve = 0.408, Guǎngyùn = 0.577, Community = 0.734. This demonstrates that the subsample was indeed very challenging, and yet two human annotators produced very similar annotations. I would like to thank Paolo Pacetto for his time and help on this task.

13 Half the time was spent annotating the 20% most challenging poems.

seems to be closer to the spirit of the original article, and shows the potential of the approach.
– The scores produced by enhancing the annotations of the original publication: re-using the concept of "set annotators", for each annotator, we collect the list of pairs of characters that have been annotated as rhyming across the entire corpus; then, we annotate odd-numbered lines based on previous rhyme judgement ("is there another character in the poem which is in rhyme position and has previously been found to rhyme with the character under consideration?"). This approximates what the original annotators would have produced, had odd-numbered lines been considered.

Table 8, Table 9, and Table 10 respectively provide the results for these three situations, indicating the precision, recall and F_1 score for the three annotators.

4.1 *General analysis*

Table 8 shows that, with the exception of the Naïve annotator which will be discussed further below, the originally published corpus produces poorer results than in Table 9 and Table 10: by ignoring odd-numbered lines, the annotations naturally cannot identify rhymes in those lines, leading to poorer recall values and consequently poorer F_1 scores.

In turn, that the scores in Table 9 are better than in Table 10 can be explained by the former being an easier exercise than the latter: indeed, in *shī* poetry, the last characters of even-numbered lines are normally always involved in a rhyme, while those of odd-numbered lines might or might not. Considering only even-numbered lines is therefore easier. This explains the very poor performance of the Naïve annotator in Table 10 compared to Table 9: according to the numbers, looking at two random characters on even-numbered lines in a random poem, the probability that they rhyme is 96% (for *shī* poetry!); once both odd- and even-numbered lines are considered, that probability falls to 36%, making the Naïve annotator wrong in 64% of its positive rhyme judgements (the '1's in our binary lists above). This means that while the Naïve annotator is not in itself a good annotation strategy in the general case, it is however a very good tool to annotate even-numbered lines (perfect recall,[14] fairly high precision) and train a Community annotator on those annotations, at least on corpora where even-numbered lines generally rhyme and poems tend to have a single rhyme throughout. If one looks at other genres of poetry, more elaborate knowledge (common rhyme patterns, for instance) could be required as a basis.

14 Because it considers everything to rhyme, it will always have a recall of 1.0, i.e. all actual rhymes are identified.

TABLE 8 Scores for the originally published corpus (missing odd-numbered lines) against the hand-annotated corpus

Annotator	Precision	Recall	F_1 score
Naïve	0.96	0.78	0.86
Guǎngyùn	1.0	0.64	0.78
Community	1.0	0.75	0.86

TABLE 9 Scores for the original corpus against the hand-annotated corpus, ignoring odd-numbered lines

Annotator	Precision	Recall	F_1 score
Naïve	0.96	1.0	0.98
Guǎngyùn	1.0	0.83	0.91
Community	1.0	0.97	0.98

TABLE 10 Scores for the enhanced original corpus (annotations for odd-numbered lines have been inferred) against the hand-annotated corpus

Annotator	Precision	Recall	F_1 score
Naïve	0.36	1.0	0.53
Guǎngyùn	0.98	0.79	0.88
Community	0.98	0.94	0.96

4.2 *Comparison of the Community and* Guǎngyùn *annotators*

Leaving the Naïve annotator aside, the three tables demonstrate that, using the F_1 score as a metric, the Community annotator is always far better than the *Guǎngyùn* annotator. This is a point that was already raised in the original article presenting the approach: throughout the Táng and the Sòng, the *Guǎngyùn* gradually lost its relevance, as a result of the pronunciation of the characters changing and poets feeling less bound to refer to the rhyme book for poetic composition.

Interestingly, both the *Guǎngyùn* and the Community annotators have very high precision—perfect, even, in the first two situations—which means that when they consider two characters to rhyme, they practically always do rhyme. This result suggests that while there may have been merges of rhyme categories (characters that the *Guǎngyùn* considers not to rhyme did actually rhyme in poems), splits seem to have been comparatively rare:[15] if splits were common and the annotators missed it (i.e. considered as rhyming the characters that are no longer rhyming), precision would drop.

If they both have similar precision, what distinguishes the two annotators is their recall: because the *Guǎngyùn* was overly prescriptive, using the *Guǎngyùn* to annotate poems produces poor results: it very often considers two characters not to rhyme even when they actually do. It is possible that, given a wider corpus of hand-annotated poems, one would find that the *Guǎngyùn* produced better results at the beginning of the Táng than at the end of the Sòng, while a Community annotator could be trained for the desired period, keeping its performance high.

4.3 *Performance as a function of annotator agreement*

As noted in the Sampling section, because of the oversampling of certain categories of poems, the scores presented above were obtained by computing scores for each annotator agreement category and then by weighting those scores to obtain a global score. A breakdown of the scores by category is presented in Table 11 for F_1 scores only. The results show that the Community annotator is expected to have near perfect performance over the entire corpus, with a F_1 score of 0.96.[16] This suggests that the approach of building an automatic annotator using community detection is a good one, at least for the *QTS* and *QSS*, and that the published annotated corpus is a reliable data source. In fact, it is worth noting that the Community annotator is as good as, or better than the *Guǎngyùn* and Naïve annotators for all categories aside from when "Community disagrees", which only accounts for 0.14% of poems.

Aside from the negligibly small "Community disagrees" category, the category where the Community annotator performs the worst (F_1=0.77) is the one where all annotators disagree,[17] followed by the one where the *Guǎngyùn* and Community annotators agree against the Naïve one. This suggests that the best

15 None seems present in the evaluation sample, but one is reported in Baley (2022b: 75).

16 A perfect score would be 1.0.

17 Should one want to improve on the published data, a focus on these poems, which represent 6.5% of the corpus, would seem to prove most useful. However, this represents nearly 16,000 poems and might take around 86 hours.

TABLE 11 F_1 scores for each annotator, according to the type of inter-annotator agreement

Category	Ratio in corpus (%)	Naïve (F_1 score)	Guǎngyùn (F_1 score)	Community (F_1 score)
All agree	63.2	0.56	0.97	0.98
Naïve disagrees	5.1	0.39	0.82	0.82
Guǎngyùn disagrees	25.1	0.52	0.67	0.99
Community disagrees	0.14	0.54	0.98	0.73
All disagree	6.5	0.37	0.54	0.77
Overall (correct sampling)	100	0.53	0.88	0.96

way to improve the annotator is to analyze the type of failures that occur in these categories and to try to identify solutions to those failures. Examining the "Community disagrees" category, of the 30 poems that were annotated by hand in this category, nearly all of them were composed between the 11th and 13th centuries, and half of them fall into one of two rhyme patterns: in 7 poems, the Community annotator fails to identify as rhyming those characters whose Late Middle Chinese reconstruction rhymes in *-an*; in 5 others, it fails to consider *ru*-tone characters as rhyming. In the section below, one poem of each failure class is presented.

4.3.1 The Community annotator fails to identify an *-an* rime

Table 12 presents the poem *Fá jí piān* 伐棘篇 (A Piece on Cutting Brambles) by Lù Zhèn 路振 (c. 957–1014) as annotated by the Community annotator, along with the pronunciation of each character in rhyme position at various stages, as reconstructed by Pulleyblank (1991). In his reconstruction, Late Middle Chinese represents the Cháng'ān 長安 dialect of the High Táng (8th century) while Early Mandarin represents the dialect of Dàdū 大都 around 1300.

In the poem, the character in rhyming position of each line is reconstructed as rhyming in *-a(ː)n* in LMC, while the rhymes for the reconstructed EMC and EM are variously *-ɛ(ː)n* and *-an*. This suggests the poem rhymes in the LMC system, but neither earlier nor later. For these characters, the Community annotator produces 4 letters ([a], [b], [c], and [d]), indicating 4 rhyme groups, but the distribution of these annotations throughout the poem does not exhibit any particular pattern; along with the similarity of those rhymes, this suggests that the poem should instead be annotated with a single rhyme annotation [a] throughout and that the Community annotator failed to identify the intention

TABLE 12 Fá jí piān 伐棘篇 (A Piece on Cutting Brambles) by Lù Zhèn 路振 (c. 957–1014) and the reconstructed pronunciation of its rhyme characters at various stages

Poem (community-annotated)	Early Middle Chinese	Late Middle Chinese	Early Mandarin
伐棘何所山之[d]巔，	tɛn	tian	tjɛn
秋風颺颺棘子[a]丹。	tan	tan	tan
折根破柢堅且[b]頑，	ŋwain / ŋwɛːn	ŋwaːn	wan′
斸夫趍趄汗污[b]顏。	ŋain / ŋɛːn	ŋjaːn	jan′
攢鋒束芒趨道[b]還，	ɣwain / ɣwɛːn	xɦwaːn	xwan′
蕫之森森繚長[c]藩。	buan	ffaːn	fan′
暮冬號風雪暗[d]天，	tʰɛn	tʰian	tʰjɛn
漏寒不鳴守犬[d]眠。	mɛn	mjian	mjɛn′
主人堂上多金[d]錢，	dzian	tsɦian	tsʰjɛn′
東陵暴客來窺[c]垣。	wuan	yan	ɥɛn′
舉手觸鋒身隕[d]顛，	tɛn	tian	tjɛn
千矛萬戟爭後[d]先。	sɛn	sian	sjɛn
襟袖結裂不可[d]揎，	swian	syan	sɥɛn
蹠破指傷流血[b]殷。	ʔəin / ʔɛːn	ʔjan	jan
神離氣沮走蹁[d]躚，	N/A	N/A	N/A
數尺之牆弗復[b]攀。	pʰain / pʰɛːn	pʰaːn	pʰan
索頭醜奴搔河[d]壖，	N/A	N/A	N/A
朔方屯師連七[d]年。	nɛn	nian	njɛn′
木波馬領沙填[d]填，	dɛn	tɦian	tʰjɛn′
氣脈不絕如喉[d]咽。	ʔɛn	ʔjian	jɛn
官軍虎怒思吼[c]軒，	xɨan	xian	xjɛn
强弩一發山河[d]穿。	tɕʰwian	tʂʰyuan	tʂʰwɛn
將不叶謀空即[a]安，	ʔan	ʔan	an
翫養小醜成隕[d]顛。	tɛn	tian	tjɛn
推芻挽粟徒喧[c]喧，	xuan	xyan	xyɛn
邊臣無心靖國[b]艱。	kəin / kɛːn	kjaːn	kjan
爲余諷此伐棘[d]篇。	pʰjian	pʰjian	pʰjɛn

of the poet. A closer look at the rhyme groups [a] through [d] and their reconstructed pronunciations shows some pattern:

– [a] rhyme characters are reconstructed with an *-an* in all periods.
– [b] rhyme characters are reconstructed as *-ɛːn* in EMC, *-aːn* in LMC and *-an* in EM.
– [c] and [d] are less clear, but [d] had either *-ian* or *-ɛn* in EMC, *-ian* or *-yan* in LMC and usually *-jɛn* in EM, while [c] had a back vowel or glide + *-an* in EMC, usually a front glide + *-an* in LMC and a front glide + *-ɛn* in EM.

These patterns show that these characters exhibit a rhyme merge from EMC to LMC and then a conditioned rhyme split from LMC to EM. The fact that one can identify such patterns suggests that the Community annotator captured the rhyming behavior of those characters throughout the Táng and the Sòng (approximately the time span between Early Middle Chinese and Early Mandarin), producing an annotator that is slightly over-prescriptive because it lacks temporal resolution and offers a global representation of the rhyming situation across six centuries. Baley (2022b: 67) also presented a case study on the *-an* rhyme, where a contrast between a Táng annotator and a Sòng annotator showed the value of training annotators for specific time periods.[18]

4.3.2 The Community annotator fails to identify the loss of *-t* and *-k* codas

The second most common category of poems in which the annotations of the Táng and Sòng corpus are wrong concerns the poems that contain characters having an entering tone in Middle Chinese, particularly the ones with *-t* and *-k* codas. Table 13 presents Yáng Shí's 楊時 (1053－1135) poem *Sòng Cài Ānlǐ* 送蔡安禮 (Sending Off Cài Ānlǐ). Leaving aside for a moment the ní 尼 rhyme — annotated as [c] — the other lines are annotated as [a] and [b] with no discernible pattern, suggesting that the poet did not distinguish these two categories. Looking at the Middle Chinese reconstructions, one sees that the two groups correspond to characters rhyming in *-k* for [a] and those rhyming in *-t* for [b], both groups having front vowels, usually *i*.

18 One can use time-bound annotators by referring to the birth and death years of the poet, but there are corner cases: for instance, of the 7 *-an* poems that the Community annotator incorrectly annotates, one is Sū Shì 蘇軾 (1037－1101) *Hé Táo "guī yuántián jū" liù shǒu: qí yī* 和陶歸園田居六首 其一 (Six Poems Following [the rhymes of] Táo [Yuānmíng's] "Returning to my Dwelling Amongst the Gardens and the Fields": 1st Poem). As the name suggests, this poem uses the rhyme of the Six Dynasties poet Táo Yuānmíng 陶淵明 (365–427) and must reflect the phonology of that time, not that of Sū Shì's.

TABLE 13 Sòng Cài Ānlǐ 送蔡安禮 (Sending Off Cài Ānlǐ) by Yáng Shí 楊時 and the reconstructed pronunciation of its rhyme characters at various stages

Poem (community-annotated)	Early Middle Chinese	Late Middle Chinese	Early Mandarin
眷言與君違，寤寐念往[a]昔。	siajk	siajk	siˇ
結歡自童稚，分比膠投[b]漆。	tsʰit	tsʰit	tsʰiˇ
乖離成參商，出沒俱齊[b]汨。	mɛjk	mjiajk	mi`
羲和鞭日御，過眼飛鳥[b]疾。	dzit	tsɦit	tsi´
五載一相逢，俯仰如昨[b]日。	ɲit	rit	ri`
論情方繾綣，念子又何[a]適。	ɕiajk	ʂiajk	ʂiˇ
行矣不可留，悁悒心若[b]失。	ɕit	ʂit	ʂiˇ
人生惟所遇，行止或使[c]尼。	ɳit	ɳit	[ni]
况復各宦遊，聚散何可[b]一。	ʔjit	ʔjit	ji`
嚶嚶黃鳥聲，上下求其[b]匹。	pʰjit	pʰjit	pʰiˇ
俛首聽遺音，飄零淚橫[a]臆。	ʔik	ʔiăk	ji`

The inter-rhyming of these characters is evidence for a loss of the distinction between *-t* and *-k* stops in Yáng Shí's dialect: Yáng Shí lived in the 11th and 12th centuries and was from the Northwestern city of Huàyīn 華陰 in modern day Shǎnxī 陝西, and beyond his poem, we know that in Northwestern dialects the *-t* coda was lost, while for *-k* codas the loss of the coda was accompanied by a diphthongization of the vowel, namely front-vowel + *-k* acquiring a *-j* off-glide and back-vowel + *-k* acquiring a *-w* off-glide, see Shen (2020:175). In the poem, all the *-t*-coda characters ([b]) had an *i* vowel, resulting in an *-i* rhyme after the loss of the coda, while all the *-k*-coda characters ([a]) had a front vowel, resulting in a front-vowel + *-j* rhyme after the loss of the coda, the front vowel eventually assimilating with the off-glide, also producing an *-i* rhyme. The complete inter-rhyming of the [a] and [b] annotations in the poem proves that these changes had already occurred in Yáng Shí's dialect at the time of composition.[19]

19 On *ní* 尼: beyond its open syllable pronunciation, Sagart (2004: 73) reports the fǎnqiè「女乙」(pointing to *ɳit* in EMC) for a quote of the Mencius to which the line of our poem is a reference ('行或使之，止或尼之' vs. the poem's '行止或使尼'). However, this pronunciation seems to have been too rare in the *QTS* and *QSS* for the community annotator to learn it, and it always groups *ní* 尼 with open syllable *-i* rhymes, explaining the [c] annotation. In any case, by the time of the poem, the coda would have been lost, *ní* 尼 then rhyming with the other *-i* open syllables.

Similarly to the previous poem, that the Community annotator failed to capture this phonological phenomenon can be explained by the annotator learning a general picture of the rhyming behaviors of poetry across the Táng and the Sòng, and this loss of codas being a relatively late phenomenon in this corpus, it was not captured. Beyond the need for a time-aware annotation scheme, this poem also points to the need for space-aware annotators: most of the poets of the Sòng corpus came from the South, which means that rhyming behaviors such as seen in Yáng Shí's poetry, showing signs of a Northwestern dialect, would have been marginal if considering the corpus as a whole.

5 Conclusion

In this paper, I have examined the proposal made in Baley (2022b) to use rhyme networks and graph community detection as a strategy to build automatic rhyme annotators for Chinese poetry. To this aim, after evaluating previously proposed metrics, I have proposed an alternative approach to evaluating the quality of the annotations produced: an F_1 score of character-pair rhyme judgements. Then, I have presented an approach to produce a sample of a corpus that is both sufficiently small for annotation by hand to be tractable and sufficiently large to provide statistically usable results. Such a sample was extracted from Baley's (2022b) published corpus before being manually re-annotated and published for others to re-use.

Then, the quality of the automatic annotations previously published was evaluated using the metric and the hand-annotated corpus: the results confirm that the Community annotator outperforms the *Guǎngyùn*-based annotator, and demonstrate that the community detection-based approach is expected to perform nearly perfectly over the corpus of 250,000 poems of the collections of *shī* poetry of the Táng and the Sòng, so that the corpus previously published is a reliable source of data for rhyming practices in those periods, thereby saving over 1300 hours of manual annotation work. Looking forward, the research community should aim to develop gold standard annotations for the entire Chinese poetic corpus: where data is sufficient, automatic annotators using rhyme community detection can be used to speed up the work, and hand-annotated samples should be produced to estimate the quality of the automatically annotated corpora. The production of such gold standard annotations will establish a strong point of reference for diachronic and synchronic analyses of Chinese phonology.

Finally, I have analyzed a class of poems for which the annotations produced the lowest scores, namely those where all of Baley (2022b) annotators produce

a different annotation. I found that over half of these poems fell into one of two classes: very late Southern Sòng poems with an *-an* rhyme and mid to late Southern Sòng poems showing evidence of loss of *-t* and *-k* codas in historically *ru*-tone characters. A careful analysis of the context of composition showed these poems to be in line with phonetic reconstructions of those characters based on rhyme books and rhyme tables. The failure to identify these phonological phenomena was explained by the fact that the Community annotator trained in the original paper had no awareness of context—neither spatial nor temporal—and produces a generic "Táng and Sòng" rhyme system. I previously showed that it was possible to exploit a finer time-granularity by training Táng- and Sòng-specific annotators, Baley (2022b: 70) producing annotators that were more accurate for poems from the period on which they were trained. I would like to conclude that this is the way forward: in the two case studies, an annotator trained on a late Sòng *shī* corpus would likely have produced the correct annotations. Further research should therefore be conducted on the development of time-aware—and possibly even space-aware—annotators. Such tools would be an important step towards the automatic—or at least assisted—study of diachronic and synchronic phonological variations in Chinese.

Supplementary materials

The hand-annotated corpus is made available in Baley (2022c) and can be used to compare the quality of different annotators on a range of Táng and Sòng *shī* poems of various rhyming complexities. The source code used for the sampling done in this article, as well as for the evaluation of the annotator, is made available at Baley (2022d); it can be used to reproduce the results found in this article or can be freely adapted to apply to other datasets.

Acknowledgments

I gratefully acknowledge the support of the Arts and Humanities Research Council ('Han Phonology: When Chinese became Chinese', AH/V008722/1) for support in the course of this research.

I would also like to thank Paolo Pacetto for his time annotating poems by hand for the purpose of inter-annotator evaluation.

References

Amigó, Enrique, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. 'A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints'. *Information Retrieval* 12 (4): 461–486. https://doi.org/10.1007/s10791-008-9066-8.

Bagga, Amit, and Breck Baldwin. 1998. 'Entity-Based Cross-Document Coreferencing Using the Vector'. In *ACL '98/COLING '98: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1:79–85. https://doi.org/10.3115/980845.980859.

Baley, Julien. 2022a. 'Automatically Annotated Quan Tang Shi and Quand Song Shi'. Zenodo. https://doi.org/10.5281/zenodo.7138623.

Baley, Julien. 2022b. 'Leveraging Graph Algorithms to Speed up the Annotation of Large Rhymed Corpora'. *Cahiers de Linguistique Asie Orientale* 51 (1): 46–80. https://doi.org/10.1163/19606028-bja10019.

Baley, Julien. 2022c. 'Hand-Annotated Sample of Tang and Song Poems for Rhyme Judgement Evaluation'. Zenodo. https://doi.org/10.5281/zenodo.7139353.

Baley, Julien. 2022d. 'Evaluating Rhyme Annotations for Large Corpora: Source Code'. OSF. https://doi.org/10.5281/zenodo.7473520.

Baxter, William H. 1992. *A Handbook of Old Chinese Phonology*. Berlin; New York: De Gruyter Mouton.

Baxter, William H., and Laurent Sagart. 2014. *Old Chinese: A New Reconstruction*. 1 edition. Oxford; New York: Oxford University Press.

Heusden, Ruben van, Jaap Kamps, and Maarten Marx. 2022. 'BCubed Revisited: Elements Like Me'. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, 127–132. ICTIR '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3539813.3545121.

Jacques, Guillaume. 2000. 'The Character 維, 惟, 唯 Ywij and the Reconstruction of the 脂 Zhi and 微 Wei Rhymes'. *Cahiers de Linguistique—Asie Orientale* 29 (2): 205–222. https://doi.org/10.3406/clao.2000.1571.

List, Johann-Mattis. 2016. 'Using Network Models to Analyze Old Chinese Rhyme Data'. *Bulletin of Chinese Linguistics* 9 (2): 218–241. https://doi.org/10.1163/2405478X-00902004.

List, Johann-Mattis. 2019. 'PoePy. A Python Library for the Quantitative Handling of Poetry'. Zenodo. https://doi.org/10.5281/zenodo.3252142.

List, Johann-Mattis, Nathan W. Hill, and Christopher J. Foster. 2019. 'Towards a Standardized Annotation of Rhyme Judgments in Chinese Historical Phonology (and Beyond)'. *Journal of Language Relationship* 17 (1–2): 26–43. https://doi.org/10.31826/jlr-2019-171-207.

Pulleyblank, Edwin G. 1991. *Lexicon of Reconstructed Pronunciation: In Early Middle*

Chinese, Late Middle Chinese, and Early Mandarin. Vancouver: University of British Columbia Press.

Quán Sòng Shī 全宋詩 (*The Complete Shī Poetry of the Sòng*). 1998. 72 vols. Beijing: Peking University Press.

Quán Táng Shī 全唐诗 (*The Complete Shī Poetry of the Táng*). 1979. 25 vols. Beijing: Zhōnghuá shūjú 中华书局.

Sagart, Laurent. 2004. 'The Chinese Names of the Four Directions'. *Journal of the American Oriental Society* 124 (1): 69. https://doi.org/10.2307/4132154.

Shen, Zhongwei. 2020. *A Phonological History of Chinese*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316476925.

Wáng Lì 王力. 2014. *Wáng Lì Quánjí 12: Shījīng Yùndú · Chǔcí Yùndú* 王力全集 12: 诗经韵读·楚辞韵读. Beijing: Zhōnghuà Shūjú 中华书局.