Multi-Dialectal Representation Learning of Sinitic Phonology

Zhibai Jia
No.2 High School of East China Normal University
jiazhibai@proton.me

Abstract

Machine learning techniques have shown their competence for representing and reasoning in symbolic systems such as language and phonology. In Sinitic Historical Phonology, notable tasks that could benefit from machine learning include the comparison of dialects and reconstruction of proto-languages systems. Motivated by this, this paper provides an approach for obtaining multi-dialectal representations of Sinitic syllables, by constructing a knowledge graph from structured phonological data, then applying the BoxE technique from knowledge base learning. We applied unsupervised clustering techniques to the obtained representations to observe that the representations capture phonemic contrast from the input dialects. Furthermore, we trained classifiers to perform inference of unobserved Middle Chinese labels, showing the representations' potential for indicating archaic, proto-language features. The representations can be used for performing completion of fragmented Sinitic phonological knowledge bases, estimating divergences between different characters, or aiding the exploration and reconstruction of archaic features.

1 Introduction

The evolution of languages in the Sinitic family created intricate correspondences and divergences in its dense dialect clusters. Investigating the dynamics of this evolution, through comparison and proto-language reconstruction, is an essential task to Sinitic Historical phonology. However, it may be costly for researchers to manually probe through the groups in search of phonological hints. Hence, it is desirable to accelerate the process with modern algorithms, specifically, representation learning.

Graph-based machine learning (Makarov et al., 2021) have gained increasing attention in recent years, due to their versatility with data with flexible structures. Especially, missing link prediction algorithms for knowledge graphs (Wang et al., 2021) (Zhu et al., 2022) can uncover a latent structure in noisy and incomplete knowledge. In the case for learning phonological representations, using graph-based learning can allow for more comprehensive integration of multi-dialectal evidence. Thus, we propose applying graph-based techniques for multi-dialectal representation learning.

We construct a knowledge graph from the multi-dialectal phonological data, by abstracting unique phonetic components and individual characters into two kinds of nodes. Then, we connect them with edges specific to the dialect type wherein the character is associated with the given component. On the constructed knowledge graph, we train the BoxE algorithm (Abboud et al., 2020), a Box Embedding Model for knowledge base completion. Finally, we evaluate the obtained representations with unsupervised and supervised clustering, as well as MLP probes based on Middle-Chinese-derived labels, to show this tool's value for Sinitic phonological investigation.

2 Background on Sinitic Languages

The analysis of Sinitic languages face a few specific challenges due to unique phonological characteristics. These characteristics define crucial details of our design.

In Sinitic languages, morphemes are primarily monosyllabic. Hence, Chinese writing binds one syllable to each of its glyphs, known as characters. A syllable in Sinitic can be decomposed into an initial, a final and a tone. (Shen, 2020) Initials refer to the consonant-like sounds at the beginning of a syllable, which include both stops (e.g. /p-/, /b-/) and fricatives (e.g. /s-/, /ʃ-/). These initials could be combined with various finals to form syllables. Finals refer to the vowel-like sounds at the end of a syllable, which included both simple vowels (e.g. /-a/, /-i/, /-u/), complex vowels (e.g. /-ai/, /-ao/, /-ei/), and vowels combined with consonant codas (/-m/,/-n/,/-ŋ/,/-p/,/-t/,/-k/). Tones refer to the pitch patterns

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop, pages 19–29

associated with syllables in Chinese. Tones could distinguish between words that were otherwise homophonous, and they were an important part of the Chinese phonological system.

Due to the early conception of the Chinese writing system, syllables from different Sinitic languages can usually be aligned to each other through a written form. As this alignment is typically implemented in databases of raw Sinitic data, the difficulty of cognate identification is drastically reduced, facilitating analysis. However, the simple syllable structure introduces large amounts of homophones, words sharing same pronunciations, into Sinitic languages. This hinders the use of the comparative method in reconstructing a Sinitic proto-language. The existence of a supersegmental tone feature also complicates a historical analysis of Sinitic languages.

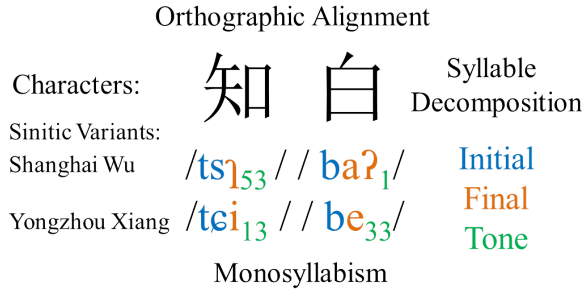

Figure 1: Highlighting key characteristics of Sinitic relevant to our approach. Characters are the central identity in the multi-dialectal representations. The orthographic alignment of sub-syllable components form the structure of data used in this study.

Two factors that motivate the use of a graph-based method include the uniform structure of Sinitic syllables and their intimate relationship with characters. The intuitive syllable decomposition and the glyph-based alignment inspire viewing the components contextualized in various dialects as different "observations" of a single character. Theoretically, these observations are derivable from the reading of the character in the proto-language.

3 Related Work

The practice of computationally-aided proto-language construction, often associated with cognate identification, has been extensively considered in the past two decades (Nerbonne et al., 2007). Examples include (Steiner et al., 2011) which draws insights from bio-informatics and the classical comparative workflow, and (List et al., 2017), which compared many methods for cognate identification. An relevant insight from the latter paper is that language-specific methods often outperform language-general ones, especially for languages like Sinitic. An epitome of neural methods for proto-language reconstruction would be (Meloni et al., 2021), in which Latin is reconstructed from Romance descendent languages with a encoder-decoder structure. Though, our approach differs from their study in many crucial aspects. In Meloni et al. 2021, the reconstruction is supervised, with the proto-language Latin provided at training time. But our method targets not only documented proto-languages like Middle Chinese, but also unknown, intermediate varieties in the development from ancient Sinitic to modern dialects, which requires an unsupervised approach. Additionally, in term of techniques, their use of GRU and attention-based transducers contrasts with our emphasis on a graph-based method.

Considering the representation learning of Sinitic, we found abundant literature on the topic of speech recognition (Ma et al., 2022), segmentation and synthesis, which often yield representations of certain phonological relevance as by-product. Though, these studies devote heavily to a few major languages, specifically Mandarin or Cantonese, and, since they are rarely claim motivation from historical phonology, seldom take a multi-lingual or multi-dialectal approach.

While speech representation learning often serve the aformentioned purposes, the proposals of using neural networks to model phonetics and phonology from either symbolic abstractions or acoustic data in order to examine theories in these fields are relevant to this study. Unsupervised binary stochastic autoencoders were explored in (Shain and Elsner, 2019). GAN (Generative Adversarial Networks) was used in (Begus, 2020). These proposals modeled perception and categorization, in relation to language acquisation. Most interestingly, representation learning has been applied for discovering phonemic tone contours in tonal languages(Li et al., 2020), of which a great portion are Sinitic Languages. However, these proposals again rarely address issues from historical phonology.

Finally, it should be noted that the concept of transforming porous data in a regular, matrix-like form to a loose, graph-like form for flexibility in processing, while essential to the designs of this paper, is not novel in the literature. Rather, it orig-

inates with the GRAPE framework in (You et al., 2020). Notably, when the data in question concerns Chinese historical phonology, it coincides with Johann-Mattis List's proposals for introducing network methods into computational linguistics and Chinese historical phonology. Generally, this line of work should be considered most relevant to our study (List, 2018; List et al., 2014; List, 2015). List (2018) approaches issues spanning character formation, Middle Chinese annotation, as well as Old Chinese reconstruction with network methods. List et al. (2014); List (2015) examines dialect evolution with display graphs, with a focus on the complex word-borrowing dynamics between the dialect families. He calls for colleagues to lend more attention to data-driven, quantitative methods. Our proposal answers List's call by bringing together knowledge graphs with Chinese historical phonology. Furthermore, the utilization of SOTA representation learning extends beyond the scope of the aforementioned work.

4 Method

The graph-based method for representing dialect data has the benefit of making the model more flexible, robust, and efficient at using porous, incomplete data. This is particularly important since investigations into dialects are often uncoordinated, resulting in a large amount of partial character entries, where only some columns have pronunciations while others are missing. It could be argued that we can use missing data imputation to alleviate the issue, and continue processing the dialect data in a matrix form, perhaps with feed-forward neural networks or denoising autoencoders(Vincent et al., 2008). However, traditional missing-data imputation techniques may create fictitious syllables that violate the phonotactics of that dialect when imputing initials or finals according to the mode of a type. Conditioning the initials or finals on each other will cause higher-order dependencies that are hard to solve. Therefore, by keeping the spaces untouched and using paired comparisons, the graph formalism circumnavigates the problem. This formulation may also allow for auxiliary input features, such as basic phonological knowledge about the nature of phonemic contrast, to be injected into the model. On this graph, we learn the embeddings with the BoxE algorithm, to be discussed below.

4.1 Construction of a Multi-Dialectal Knowledge Graph

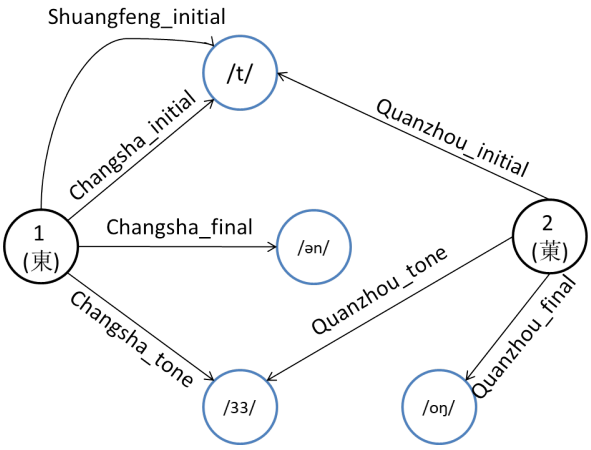

Figure 2: Partial Illustration of the Phonology Knowledge Graph. The numerals represent the indices representing the Chinese characters and the glyphs for what they represent. /33/ is a tone in Chao's notation. The other nodes are segments represented in the International Phonetic Alphabet. The text labels for the edges demonstrate the how edges are categorized according to both dialect and phone type. Note that it is bi-partite by nature, as edges can only occur between "phonemic" nodes and "character" nodes, colored blue and black in the figure.(This is not provided explicitly)

We expressed the data with a knowledge graph and trained the representations through an auxiliary task of completing the multi-dialectal knowledge graph. With a graph-based technique, the representations can be more robust to noisy and porous data. Additionally, the method will be more flexible, allowing for auxiliary input features to be injected.

We construct a graph by leveraging the characters, as well as individual initials, finals and tones from various dialects as nodes. (See Figure 2) .For instance, the fact of character C having an initial I in dialect D is modeled with an edge from C to I. The edge has type specific to the dialect D and the category of the component, which is an initial. This edge type can be denoted as "D-initial". Demonstrated in Fig. 2, C could be character No. 1, when I is /t/ and the edge is "Changsha_initial".

After constructing the graph, character-level and component-level representations are trained simultaneously. The knowledge graph algorithm attempts to model the nodes features as well as a prediction function so that, when given a character node and a type of link, the corresponding pronunciation node can be predicted with maximum likelihood. In this process, the model implicitly gen-

erates hypotheses about character pronunciations missing or unseen in training, as well as historical relationships between the syllables.

If there are M characters with readings from N dialects involved in an experiment, the upper bound for the number of edge types will be $3N$. Assuming that $F_1 + F_2 + F_3$ unique initials, finals and tones could be found within the aggregated phonological systems of the N dialects, the upper bound for number of nodes is $M + F_1 + F_2 + F_3$. The graph size scales sub-linearly with the number of dialects, since as more dialects are considered, their phonemic inventories will start to overlap and exhaust.

Following convention in knowledge base research, the graph is presented in Triples of Head-Relation-Tail format.

4.2 The Box Embedding Model

In pilot tests, We considered various algorithms from the field of graph representation learning and knowledge base completion for application. In the process, it is revealed that few algorithms are inherently suitable, as there are many subtle requirements in this context:

1. Models designed for knowledge graphs are more suited to this application than general graph learning algorithms, since the graph to be processed is heterogeneous, besides carrying edge type as information.

2. The model must have strong capacity for modeling multiple unique relations between the same two nodes. It is very common for one character to have the same initial across different dialects. This rules out many translation-based models, that, when given different relations, always predict different tail nodes. Prominent examples of such models include TransE (Bordes et al., 2013) and RotatE (Sun et al., 2019).

3. If the model uses inverse triples as an augmentation technique, then the model should also be expressive in many-to-one and one-to-many relations, because one initial or final will be mapped to numerous characters.

4. Of the applicable algorithms, interpretability should be prioritized, since we hope to extract interpretable phonological knowledge from the obtained representations. This casts doubt on a another large family of knowledge graph models, namely the bi-linear models, epitomized by RESCAL(Nickel et al.) and DistMult(Yang et al., 2015).

After consideration, we chose BoxE for its expressiveness and tolerance to many-to-one relationships, due to its Box embedding designs. Empirically, we also demonstrate that the BoxE is relatively optimal for the phonological task through comparison with RotatE (Sun et al., 2019) and ComplEx (Trouillon et al., 2016) in Table 4.

Here is a brief description of the BoxE algorithm. It is a translational model that embeds each node with two vectors: e_i, which represents the position vector, and $b_i \in \mathbf{R}^d$, which represents the translational bump. These vectors are obtained after incorporating triples into the model. Additionally, each edge type is defined with two hyper-rectangles $r^{(1)}$ and $r^{(2)} \in \mathbf{R}^d$. To satisfy the relation R between entity E_1 and E_2, there is $e_1 + b_2 \in r^{(1)}$ and $e_2 + b_1 \in r^{(2)}$. Intuitively, this means that E_1 and E_2 "bump" each other in hyperspace $\mathbf{R}^d$ by some distance. If the new vectors fall within the bounds of the associated boxes, then the proposition is considered probable. To facilitate gradient descent, the boxes have relaxed borders. It is worth noting that BoxE is also capable of hyper-graph learning as it accepts higher arity relations as input, though we did not exploit this feature for this study.

Our training objective was to maximize the score or probability of given relations. To elaborate, this means maximizing the chance of predicting masked initials/finals/tones of some character in some dialect with the unmasked components associated with that character, from both within and without the dialect. This is analog to the comparative method in Historical Phonology, as the model implicitly reconstructs a latent "proto-language", from which the descendent languages can be deduced (or, "decoded") with maximum likelihood.

5 Data and Experimental Setup

We use pronunciation data from four varieties of Xiang Chinese Changsha 長沙, Shuangfeng 雙峰, Guanyang Wenshi 灌陽文市, and Quanzhou Xiancheng 全州縣城., spoken primarily in Hunan Province, provided by CCR(Huang et al., 2011), and retrived with Comparative analysis toolset for Chinese dialects(Huang, 2021). We also obtain labels of Middle Chinese readings from the same source. In this work, Middle Chinese refers to

the phonological system recorded in the dictionary Qieyun, from the year 601 AD. It was supplemented in the Song Dynasty into the dictionary Guangyun, from which this study draws data. Middle Chinese is literary and may not reflect the colloquial speech of China in any time or place. However, most phonological systems of modern Sinitic languages (with the notable exception of the Min Languages) can be derived from the Qieyun system. Thus we treat it as a useful protolanguge model for most Sinitic Languages.

We operate on symbolic abstractions instead of raw acoustic data, as all the data have been transcribed into IPA in the database. One row of data corresponds to readings of one Chinese character. Internally, each character is mapped to a unique identifier, which is the character's serial number in Guangyun. For every variety of Chinese, there are four columns, corresponding to initial value, final value, tonal value and tonal type of a given character's pronuciation. The tone type argument is actually redundant, and it is assigned manually by investigators. In each dialect, there is a one-to-one correspondence between one tone value with one tone type. Between two dialects, tones arising from the same Middle Chinese tone are given same names. Hence, the tone type feature introduces prior expert knowledge about the historical origin of tones. However, we expect the model to derive the historical tones without any diachronic expert knowledge. Hence, we discard the tone type feature, and use only the three values for this study.

5.1 Processing of Duplicate Data

Characters in Sinitic can be polyphonic, that is, sometimes a character will be mapped to multiple readings in one dialect. This results in duplicate data in the dataset. For convenience, we drop the extra pronunciations and keep only the first line for every entry. Though, there can be ambiguity surrounding the correspondence of readings for polyphonic characters. For instance, the first reading entry for a polyphonic character in dialect A might be cognate with the second reading entry for the character in dialect B. However, our naïve approach will match all the first entries to each other. Additionally, two dialects may inherit only partial readings of a polyphonic character in the proto-language. Hence, this procedure potentially introduces erroneous alignment into the model.

5.2 Split of Training, Testing and Validating Datasets

The model was not trained with all the data, so as to examine the robustness of the model. Instead, some triples are diverted to form testing and validating datasets. Unfortunately, assignment in this context is slightly more complicated than simple stochastic choice. There is the scenario where all initial (final/tonal) information about one character is diverted from training. In this case, the model will not be able to correctly embed this character. To circumvent this issue, we mandate that at least one feature from any of the three compositional types is retained in the training set for any character. In the four Xiangyu in this case, the result is empirically a split of 80.50%:12.52%:6.98%.

5.3 Data Statistics

The initials, finals and tones count for the four dialects are listed in Table 1. A total of 2805 characters is included, but not every character has the corresponding phonological data documented in every dialect. In the training set, there are 22300 entries.

5.4 Model Setup

For the parametric size of the model, see Table 2. We employ the BoxE algorithm implemented in the Python library PyKeen (Ali et al., 2021b,a). We did not fine-tune the model or any model parameters, so as to demonstrate the capability of the model in even in a highly suboptimal setting.

	Initials	Finals	Tones
Changsha	21	38	11
Shuangfeng	28	35	11
Guanyang	28	42	5
Quanzhou	26	43	4

Table 1: Data Statistics

Parameter	Value
Vector and hyperbox dimension	64
Number of nodes	2946
Number of edge types	12
Cumulative parameter size	378624
Optimization algorithm	Adam
Number of epochs	2000

Table 2: Model Parameters

6 Experimental Evaluation

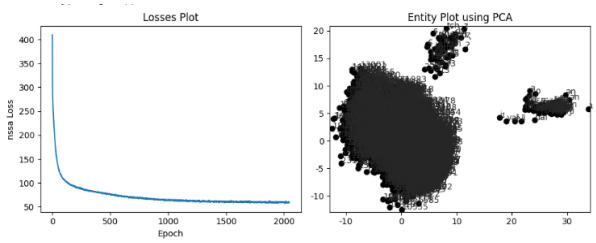

Figure 3: Preliminary Visualization of Training Dynamics and Trained Embeddings.

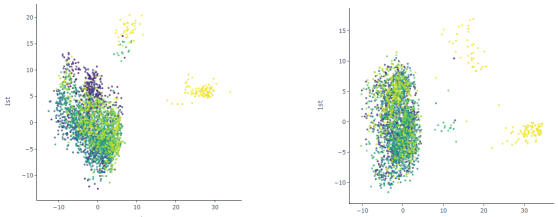

Figure 4: UMAP(McInnes et al., 2018, McInnes et al., 2018,Uniform Manifold Approximation and Projection) decomposed visualizations of the translational bumps (a) and position embeddings (b). The coloring reflects a point's index in the Guangyun, which is sorted according to rhyme.

6.1 Canonical Evaluation of Model

The convergence of the model, and a preview of the spatial distribution of embeddings can be seen in Figure 3. The model quickly converges. The entity plot decomposed with PCA reveals a mass of character readings "ejecting" two groups of entities, respectively the combination of all initials and tones, and all finals, which is in accordance with the bi-partite and heterogeneous nature of this graph.

Canonically, BoxE is evaluated with the hit@n metric and MRR (mean reciprocal rank) for link prediction. On the validation set, our model achieved hit@1:51.25%, hit@5: 87.19%, hit@10: 93.76% on the "tail" batches. The head batches are not relevant because they involve "predicting characters from initials/finals", of which there is many to one. In Table 4, we demonstrate empirically the superiority of the BoxE algorithm over other common knowledge graph algorithms on this phonological task. A clearer visualization of the embedded points can be seen in Figure 4. Guangyun ensures that rhyming characters (having the same final) have similar coloring on the map. The coloring is only a reflection of the point's serial in the dataset and does not have any quantitative interpretation. Presumably, the translational bump for characters will contain more relevant information to historical phonology, as they designate which component types to "bump into the box." Without mention, all experiments are carried out on the bump embeddings and not positions. However, empirically we find that the two kinds of embeddings are interchangeable.

6.2 Examining Contrastive Information

In this section, unsupervised clustering is used to evaluate contrastive information in the embeddings. Based on the hypothesis that the phonological structures of the dialects are co-embedded in the latent structure of embeddings, we determined if the high-dimensional embeddings retain information associated with the theoretic categories of the input dialects, a similar task to Tilsen et al. 2021. After applying a clustering algorithm to the embedded characters, the information yield [1] of the found categories against input categories of initials, finals and tones is computed. A higher information yield indicates that the clusters found by unsupervised clustering were more interpretable with respect to the input phonemic categories. [2] [3]

The clustering algorithms used for dissecting the cloud of embedded characters include HDBSCAN (McInnes and Healy, 2017,A density based method), Affinity Propagation, K-means and Agglomerated Clustering.[4] The results can be seen in Figure 5.

Affinity propagation and HDBSCAN achieved best effects on finding interpretable clusters from the datasets. Though, we find that HDBSCAN is very sensitive to the two parameters: its effect degrades when we allow for smaller clusters but demands greater confidence on the classification. Notably, HDBSCAN achieved an effect similar to affinity propogation with just 29 clusters, while the latter used 130.

The large information yields reflect that the unsu-

[1] Entropy subtracted by conditional entropy, or an empirical estimate of mutual information.

[2] HDBSCAN sometimes refuses to classify points it is not sure of. These points are combined into one category for the aforementioned purpose.

[3] Before using HDBSCAN, UMAP was first used to reduce the 64 embedding dimensions to 8 dimensions, with the neighbour parameter set to 50. This is an advised practice from the HDBSCAN documentation.

[4] The numerous methods were tried sequentially as we do not know which algorithm best recovers the latent structure of representations in accordance with theoretic categories.

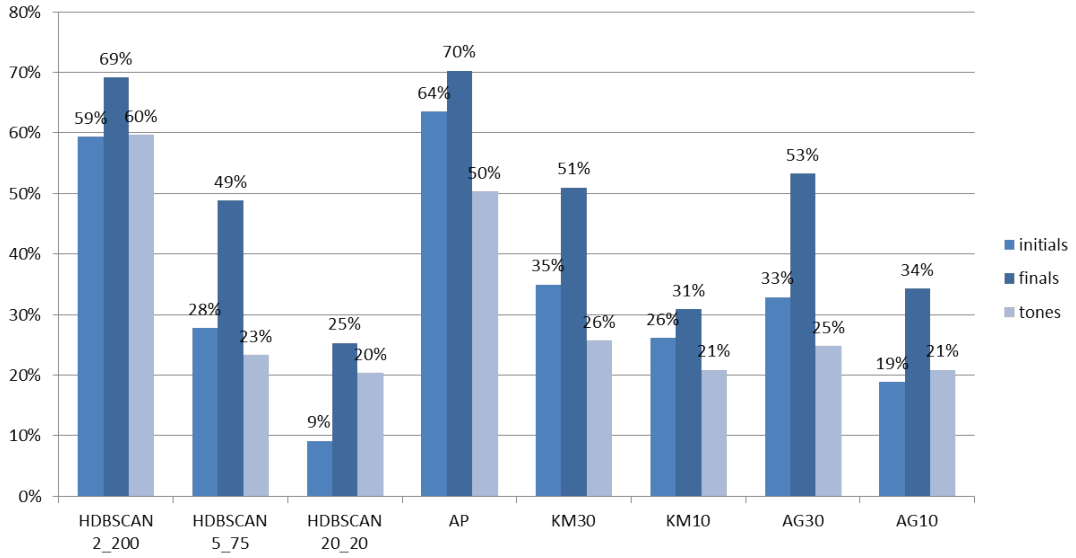

Figure 5: Information yield in percentage averaged across four dialects. For HDBSCAN, the min samples and min cluster size parameters were set to 2 and 200, 5 and 75, 20 and 20 respectively. The other three methods were employed on the original embeddings. For K-means and agglomerative clustering, the number of clusters was specified to be 30 and 10.

pervised algorithms do tend to dissect the character set along latent lines corresponding to phonological opposition in the input dialects, as shown in a partial observation in Table 3. It appears that the distribution of finals in dialects had more influence on the latent structure than initials or tones. Simply put, the characters within each unsupervised cluster are more likely to rhyme than alliterate, though both cases occur in observation of the HDBSCAN Clusters.

There are limitations to this experiment though, which will be discussed below.

6.3 Inference of Proto-language Features

In this section, we investigate the quality of our embeddings with respect to proto-language reconstruction tasks, as an important potential application of this method lies with such work. Hence, we trained classifiers in attempt to infer labels from Middle Chinese, which likely predates proto-Xiang, therefore an accessible surrogate for that proto-language.

The features to infer are Grades (等地), Voice(清濁), Tones(聲調), She (攝, a coarse division of finals), Initials (字母), and Mu(韻目,a fine division of finals).

Grades are believed to be associated with medials, a component in the front of the final (amalgamated with final in Xiangyu data). Voice is a division based on properties of the initial, in which voiced consonants, voiceless unaspirated consonants, voiceless aspirated consonants and nasal consonants are distinguished. For tones, in Middle Chinese, there were four: level, rising, departing, and entering. Of these categorical labels, there are respectively 4, 4, 4, 16, 36 and 206 unique classes. [5]

For this experiment, a train-test split of 0.67-0.33 was instated. Since phonological evolution is quite regular and systematic, we should expect decent results without a great proportion of data used for training. Accuracies below are for the test set. These values are consistently higher than a naïve baseline of guessing the mode of each distribution, proving that proto-language related features were preserved in the retrived embeddings. (See Table 5.)

The MLP generally outperforms Ridge Classification on inference for these characters, with the sole exception of tones, where RC outperforms MLP by 1.1%. The best results are attained for tones and voice, showing these features to be phonologically well preserved from Middle Chinese to Xiang languages.

Interesting observations can be drawn from the confusion matrices generated with such classification. Presumably, these matrices can offer insight

[5] Canonically so, but there are a few erroneous entries in the data we used, resulting in sometimes one or two extra categories containing a few characters. They were kept.

ID	Changsha	Shuangfeng	Guanyang	Quanzhou
0	Initial:/m/	Initial:/m/	Initial:/m/	Initial:/m/
1	Initial:/p^h/	Initial:/p^h/	Initial:/p^h/	Initial:/p^h/
2	Final:/ĩn/	Final:/ĩ/	Final:/ iɛ̃/	Final:/ ieŋ/
7	Final:/(u)ei/	Final:/ui/	Final:/ uɛi/	Final:/uei/

Table 3: Analysis of Selected HDBSCAN Clusters. In these clusters, characters are predominantly, but not exclusively associated with the listed features.

Alg. (Metric %)	Hit@1	Hit@5	Hit@10
BoxE	**51.25**	**87.19**	**93.76**
RotatE	33.11	57.47	66.18
ComplEx	9.40	24.65	35.37

Table 4: An empirical demonstration of the superiority of the BoxE algorithm for the phonological investigation task among common missing link prediction methods. The models were set to the same embedding dimension. None of the models were fine-tuned or ran for more than a single time, hence all readings should be seen as sub-optimal.

into what categories were blended, which oppositions were lost during the development of some language family. One such example is demonstrated in Figure 6. It could be seen that there is large confusion between the Xian 咸, Dang 宕 and Shan 山 Shes, and also between Xie 蟹 and Zhi 止 Shes. [6] This could indicate that in Proto-Xiang, there is confusion between these categories relative to Middle Chinese.

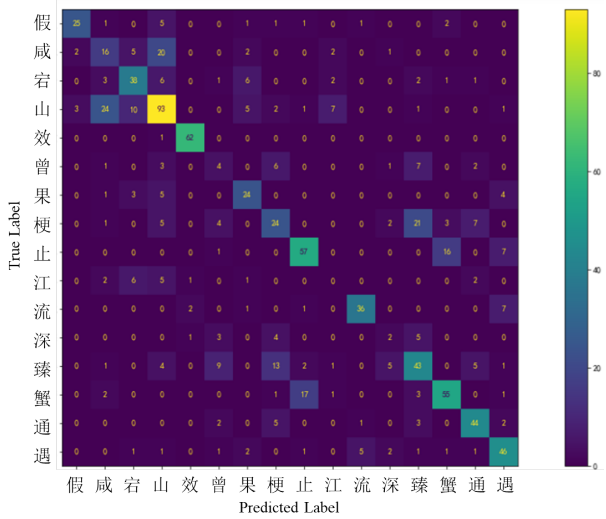

Figure 6: Confusion matrix for She.

[6] In Baxter's transcription, 咸 = *-eam*, 宕 = *-ang*, 山 = *-ean*; 蟹 = *-ea*, 止 = *-i* (Baxter and Sagart, 2014). There are only hypothetical IPA values available for these archaic categories.

7 Discussions

Our current setting only operates on pre-abstracted symbols and lacks incorporation of acoustic or articulatory evidence. Incorporating multi-modal data into a knowledge graph framework could enhance the quality of embeddings and enable more accurate representations of phonological features. Alsp, the proposed method uses shared embeddings for symbolic components across different dialects, which cannot fully capture dialect-specific variations. Investigating contextualized or dialect-specific component embeddings could improve the model's ability to capture finer-grained phonological distinctions. Finally, phonetically similar components are currently treated as independent items, which is too absolute an assumption. However, it is also possible for phonetic cues to override the correct phonological alignment in the model. In many cases, phonetic similarity does not imply diachronic homology. Two phonetically equivalent syllables from two different dialects may have different origins. Conversely, two phonetically distinct syllables from two different dialects may be cognate. The subtle balance between "phonetic" and "phonological" proximity requires further discussion.

Several lines of research may benefit from robust multi-dialectal representations. In dialectology, there is need for estimating divergence between phonological systems. That includes the divergences between its constituents, such as individual characters, phonemes and syllables. With multi-dialectal representations, this divergence can be estimated quantitatively. In historical phonology, the reconstruction of a proto-language demands deep scrutiny of dialect systems whose efficiency can be improved with manipulating the representations. Also, they can be used for completion of the phonological knowledge base. Often knowledge bases for Sinitic phonology are fragmented, due to imperfect surveys and heterogeneity of sources, etc. The representations can be used to infer missing

Algorithm(Acc %)	Grades	Voice	Tones	She	Initials	Mu
Ridge Classification	65.3	76.4	**84.1**	54.6	49.4	18.6
MLP	**70.5**	**81.1**	83.0	**61.4**	**53.2**	**26.9**
Naïve Baseline	48.4	35.4	35.6	15.3	8.1	1.8

Table 5: Comparison of Ridge and MLP probes for proto-language Feature Inference. The baseline is the accuracy obtained by uniformly guessing the most frequent class for each character.

pronunciations in different dialects to improve the quality of observations.

The graph-based method proposed in this paper benefits from phonological characteristics specific to Sinitic languages, but is also limited by these characteristics. Specifically, the process of constructing a phonological graph from words, as proposed in this study, is less natural in languages where words typically have many syllables, and vary in the number of syllables contained. In these languages, the temporal interaction of syllables within a word is a new phenomena that the graph-based method needs to adapt to. Additionally, in these languages, it will be less straightforward to tokenize the words into expressive sub-words to use as nodes in the graph. Presumably, in non-Sinitic languages, the proposed method will be most performant in other languages of the Southeast Asian Sprachbund, such as those in the Hmong-Mien or Austroasiatic families. These languages share phonological features with Sinitic languages that enable our method. On the other hand, this method will likely meet more complications outside of the local sprachbund.

8 Conclusion

This paper demonstrated the potential of graph-based representation learning in Chinese Historical Phonology. The representations are potent in many ways, i.e. facilitating the reconstruction of minor proto-languages.

In the future, more sophisticated techniques such as deep learning models could be explored to further improve the quality of the obtained representations. Furthermore, the proposed method can be integrated with other linguistic resources, such as recordings, articulatory time series, or orthographic corpora, to enrich the knowledge base and improve the accuracy of reconstructions. With the development of modern, massive linguistic datasets such as Nk2028(nk2028, 2020), CogNet(Batsuren et al., 2022) or MorphyNet(Batsuren et al., 2021) as well as improvements in large pre-trained models, we can expect foundational models that possess emergent and meta-generalizing capabilities to arise in historical phonology or morphology. This avenue of research holds great promise for advancing our understanding of the phonology and evolution of Sinitic languages, and potentially other language families as well.

Limitations

This study stems from a novel idea for Chinese Historical Phonology Studies. As few direct predecessors could offer hindsight, there are quite a few limitations to this study that may be addressed with further work.

1. While the initial-final-tone decomposition is convenient in this context, it also limits the transferrability of the proposed tool to languages outside of the Sinosphere. This calls for further exploration of more generalizeable approaches to phonological representation learning.

2. Polyphonic characters were not fully utilized in the study, and their alignment per-reading and tokenization into separate identifiers should be considered in future work.

3. Finally, making full use of the dataset is crucial, and the stochastic train-test split used in this study may leave out important hints. Alternative sampling strategies, such as cross-validation or bootstrapping, could enhance the robustness of the results.

Acknowledgements

We are grateful for the valuable advice and feedback we received from various peers during the course of this work. Without their contributions, this research would not have been possible.

References

Ralph Abboud, Ismail Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. 2020. BoxE: A Box Em-

bedding Model for Knowledge Base Completion. In *Advances in Neural Information Processing Systems*, volume 33, pages 9649–9661. Curran Associates, Inc.

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Mikhail Galkin, Sahand Sharifzadeh, Asja Fischer, Volker Tresp, and Jens Lehmann. 2021a. Bringing Light Into the Dark: A Large-scale Evaluation of Knowledge Graph Embedding Models under a Unified Framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. 2021b. PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research*, 22(82):1–6.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: a Large Multilingual Database of Derivational and Inflectional Morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2022. A large and evolving cognate database. *Language Resources and Evaluation*, 56(1):165–189.

William H. Baxter and Laurent Sagart. 2014. Old chinese: A new reconstruction.

Gasper Begus. 2020. Modeing unsupervised phonetic and phonological learning in Generative Adversarial Phonology. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 38–48, New York, New York. Association for Computational Linguistics.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Rongpei Huang, Xiufang Yang, and Daan He. 2011. Chinese Character Readings. `https://xiaoxue.iis.sinica.edu.tw/ccr/#`. Retrieved March 26, 2023.

Yihua Huang. 2021. Comparative Analysis Toolset for Chinese Dialects. `https://github.com/lernanto/sinetym`. Retrieved March 26, 2023.

Bai Li, Jing Yi Xie, and Frank Rudzicz. 2020. Representation Learning for Discovering Phonemic Tone Contours. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 217–223, Online. Association for Computational Linguistics.

Johann-Mattis List. 2015. Network perspectives on chinese dialect history. *Bulletin of Chinese linguistics*, 8:27–47.

Johann-Mattis List. 2018. More on network approaches in historical chinese phonology ().

Johann-Mattis List, Simon J. Greenhill, and Russell D. Gray. 2017. The Potential of Automatic Word Comparison for Historical Linguistics. *PLOS ONE*, 12(1):e0170046.

Johann-Mattis List, Nelson-Sathi Shijulal, William F. Martin, and Hans Geisler. 2014. Using phylogenetic networks to model chinese dialect history.

Han Ma, Roubing Tang, Yi Zhang, and Qiaoling Zhang. 2022. Survey on speech recognition. *Computer Systems and Applications*, 31(1):1–10.

Ilya Makarov, Dmitrii Kiselev, Nikita Nikitinsky, and Lovro ubelj. 2021. Survey on graph embeddings and their applications to machine learning problems on graphs. *PeerJ Computer Science*, 7.

L. McInnes, J. Healy, and J. Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*.

Leland McInnes and John Healy. 2017. Accelerated hierarchical density based clustering. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*, pages 33–42. IEEE.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4460–4473, Online. Association for Computational Linguistics.

John Nerbonne, T. Mark Ellison, and Grzegorz Kondrak. 2007. Computing and historical phonology. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology on - SigMorPhon '07*, pages 1–5, Prauge, Czech Republic. Association for Computational Linguistics.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A Three-Way Model for Collective Learning on Multi-Relational Data.

nk2028. 2020. Qieyun-js. `https://github.com/nk2028`. Retrieved March 26, 2023.

Cory Shain and Micha Elsner. 2019. Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders. In *Proceedings of the 2019 Conference of the North*, pages 69–85, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhongwei Shen. 2020. A phonological history of chinese.

Lydia Steiner, Michael Cysouw, and Peter Stadler. 2011. A Pipeline for Computational Historical Linguistics. *Language Dynamics and Change*, 1(1):89–127.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. ArXiv:1902.10197 [cs, stat].

Sam Tilsen, Seung-Eun Kim, and Claire Wang. 2021. Localizing category-related information in speech with multi-scale analyses. *PLOS ONE*, 16(10):e0258178.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. ArXiv:1606.06357 [cs, stat].

Pascal Vincent, H. Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*.

Meihong Wang, Linling Qiu, and Xiaoli Wang. 2021. A survey on knowledge graph embeddings for link prediction. *Symmetry*, 13:485.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. ArXiv:1412.6575 [cs].

Jiaxuan You, Xiaobai Ma, Daisy Ding, Mykel Kochenderfer, and Jure Leskovec. 2020. Handling missing data with graph representation learning. *NeurIPS*.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-modal knowledge graph construction and application: A survey. *ArXiv*, abs/2202.05786.