Restoring Mycenaean Linear B 'A&B' series tablets using supervised and transfer learning

Katerina Papavassileiou and **Dimitrios I. Kosmopoulos**
Postdoc researcher and Associate Professor
Department of History and Archaeology, Computer Engineering and Informatics Department
University of Patras
Patra, Greece
cpapavas@upatras.gr, dkosmo@upatras.gr

Abstract

We investigate the problem of restoring Mycenaean linear B clay tablets, dating from about 1400 B.C. to roughly 1200 B.C., by using text infilling methods based on machine learning models. Our goals here are: first to try to improve the results of the methods used in the related literature by focusing on the characteristics of the Mycenaean Linear B writing system (series D), second to examine the same problem for the first time on series A&B and finally to investigate transfer learning using series D as source and the smaller series A&B as target. Our results show promising results in the supervised learning tasks, while further investigation is needed to better exploit the merits of transfer learning.

1 Introduction

For many years the language attributed by the Linear B script (c. 1400-1200 B.C.) was a point of contention among scientists, who argued over the origin of the Mycenaean syllabary. The answer was given in 1952 by Michael Ventris (Chadwick, 1990; Fox, 2013) who, together with the philologist John Chadwick, proved that the syllables of the Linear B script form words of the Greek language and that the Mycenaean world was both linguistically and culturally linked to ancient Greece.

Mycenaean Linear B is a syllabic script. It includes syllables as well as logograms or ideograms. In summary, Linear B is structured by groups of phonetic symbols, which are accompanied by ideograms. The surviving tablets typically refer to human names, place names, agricultural production, land ownership, religious offerings, or military equipment. The Mycenaean inscriptions have been classified based on their place of origin, but also based on the category to which they belong. The place of origin is indicated by the following abbreviations: KN (Knossos), PY (Pylos), MY (Mycenae), TH (Thebes), TI (Tiryns), KH (Khania), MI (Midea), etc. The classification into categories was based on the ideograms of the tablets: Series A&B, (lists of personnel), series C (animal records), series D (sheep records), series E (grain records), series F&G (records of oil, agricultural products and their offerings), series L (textile records), etc. (Kober, 1945), (Kober, 1946), (Kober, 1948), (Ruijgh, 1977).

The main challenge faced by those dealing with the study and restoration of the Mycenaean Linear B texts, either manually (Ventris and Chadwick, 1953, 1956; Killen, 1964; Doria, 1965; Ventris et al., 1988; Meissner, 2001; Robinson and Eisenman, 2002; Pope, 2008; Duhoux and Davies, 2008; Fox, 2013; Ventris and Chadwick, 2015; Freo and Perna, 2019; Bernabé and Luján, 2020) or computationally (Papavassiliou et al., 2020; Papavassileiou et al., 2023), is the scarcity of data. Furthermore, we have to take into account the particularities presented by the Mycenaean inscriptions: a) Their eminently administrative content, b) their subject as they deal extensively with people and places and c) their state of preservation since most of them are broken, worn out or burnt. These make the infilling task very challenging.

This article contributes by investigating Transfer Learning (TL) techniques to alleviate the above mentioned data scarcity. TL is the process of taking a model that has been trained to do one task (pre-trained model) and fine-tuning it to work on a related (different or similar) task. In Natural Language Processing (NLP), pre-trained models are often used as the starting point for a wide range of NLP tasks, such as language translation, sentiment analysis, and text summarization. By using a pre-trained model, we can save time and resources, as they don't have to train a model from scratch on a large dataset (Devlin et al., 2019; Radford et al., 2019; Peters et al., 2018; Liu et al., 2019).

In this work we deal with series A&B and series D to find out the effect of TL from a series to

Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), pages 115–129

another one. To this end we have chosen to investigate the effect of TL from the series D, since it is the largest with the most tablets and resulting sequences, to the series A&B, as opposed to training the series as a whole.

We also contribute by investigating whether there is an improvement in the predictions of models learned from series D (Papavassileiou et al., 2023; Papavassiliou et al., 2020) if we remove the ideograms from the sequences. This choice is based on the observation that in Linear B, typically there is consistent separation of phonographic (syllabograms) and non-phonographic (ideograms as well as signs for measurement units and numbers) graphemes (Petrakis, 2017).

Finally we apply generative methods for the infilling problem of series A&B. Following the same tactic with series D, we create a dataset of sequences derived from the Mycenaean tablets of series A&B, excluding ideograms, and use it to learn a generative model in order to predict damaged parts of this series' tablets.

The rest of the paper is organized as follows: In Section 2 we provide an overview of previous research and studies on various methods of restoring ancient inscriptions. In Section 3 we present the dataset for our experiments. Section 4 demonstrates supervised learning for language modelling and for missing symbol recovery. Section 5 presents transfer learning techniques for infilling series A&B. Finally, in Section 6 we present our conclusions and the future work.

2 Related work

The problem of text restoration through infilling is attracting more and more attention from researchers community; however most recent results from the Natural Language Processing (NLP) community have been only partially applied, obviously due to the lack of sufficient data.

Some of the simplest models are the n-grams. These are probabilistic models for predicting the next item in a sequence of n elements and can be used to model almost any type of sequential data. They have been used for machine translation (Wolk and Marasek, 2014), but also for textual restoration. (Rao et al., 2009) and (Yadav et al., 2010) use n-gram Markov chains for texts in the Indus script. The benefits of n-gram models are their simplicity and scalability. With larger n, a model can store more context, enabling small experiments to scale up. However, when n increases, the number of possible n-grams increases exponentially and therefore the out-of-vocabulary n-grams increase as well and actually undermine the performance of the model. Obviously, the n-grams are not appropriate for long sequences.

(Roued-Cunliffe, 2010) uses a decision support system called DUGA for reading ancient documents in the Latin language found in Vindolanda (Britain). She uses the so-called cruciverbalistic approach: it begins by establishing the letters that are legible and uses them as a foundation for a subsequent hypotheses. A knowledge-base of previously interpreted documents from the same period is used to extract word lists and frequencies. These are then used to suggest different interpretations of words and letters, as well as missing parts, using a hierarchical approach from individual symbols to whole sentences. The system is therefore largely based on the experts' decisions. (Kang et al., 2021) present a multi-task learning approach based on the Transformer networks to effectively restore and translate ancient historical documents based on a self-attention mechanism, specifically utilizing two Korean historical records, one of the most voluminous historical records in the world. This work combines 3 different studies: the restoration of damaged documents (recovering), neural machine translation (translating), and the analysis of historical records (mining). The proposed model consists of embedding and output layers for Hanja and Korean, and three Transformer modules: the shared encoder (for both the restoration and translation tasks), the restoration encoder (for the restoration task), and the translation decoder (for translating Hanja sentences into modern Korean sentences). However, a large-scale training corpus is required.

Similar to our work is the PYTHIA system (Assael et al., 2019) and its follow-up system Ithaca (Assael et al., 2022). It aims to fill the missing symbols (characters) in ancient Greek inscriptions. The authors use a sequence-to-sequence framework (Sutskever et al., 2014) with Long Short-Term Memory (LSTM) networks in the encoder and the decoder. The encoder involves the input character embeddings sequence with missing characters, and a separate stream is also modelled using the word sequence as embeddings as well; an attention layer is also used. The decoder is trained to output the missing characters. They use a dataset that results from processing the epigraphical corpora of the Packard Humanities Institute (Packard Human-

ities Institute, 2005), the PHI-ML. As (Shen et al., 2020) argue in their ancient text restoration experiment, (Assael et al., 2019) perform restoration at the character-level where the number of characters to recover is assumed to be known and indicated by a corresponding number of '?' symbols. In reality, when epigraphists restore a deteriorated document, the length of the lost fragment is unknown and needs to be guessed as a first step. BLM, in essence a variant of BLM, the L-BLM, can bypass this limitation and flexibly generate completions without this additional knowledge. A single token, sized equal to the number of '?' symbols, is defined and the L-BLM is trained to predict a character to fill in and the length of the new blank to its left. Compared to our work, the problem presented by the authors of these articles (Assael et al., 2019; Shen et al., 2020) is similar in the sense that it concerns a known script and known language and uses a machine learning architecture. However, our task is more challenging, due to the fact that the corpus is of much smaller size (over 40000 inscriptions available in the aforementioned articles versus 1100 inscriptions in ours); that impedes training.

(Fetaya et al., 2020) use recurrent neural networks (LSTM) to restore fragmentary Babylonian texts. These involve ancient texts in the Akkadian language, which belong to the Semitic language family. Comparisons to simple 2-gram baseline approach (considering the previous and the next word) are made, resulting in better performance. The experiments use a dataset of 3000 transliterated archival documents belonging to economic, juridical and administrative genres. Similarly to this work, (Lazar et al., 2021) also introduce BERT-based models aiming to solve the task of predicting missing signs in Akkadian texts. The difference with the previous article (Fetaya et al., 2020) is that the completion of missing signs is done by combining large-scale multilingual pretraining with Akkadian language finetuning. Although (Fetaya et al., 2020) have small-scale data at their disposal to train the learning algorithm (c. 3000 Babylonian transliterated texts, 539-331 B.C.E.), what is emphasized by the authors is that the late Babylonian texts are structured official bureaucratic documents, e.g., legal proceedings, receipts, promissory notes, contracts and so on. This is in stark contrast to Linear B tablets, which are significantly impeded by syntactic inconsistencies. This is proved by the fact that the pre-processing of each Mycenaean tablet requires special handling, so as to extract valid sequences of Mycenaean words in accordance with the principles of Mycenaean language. Another BERT-based model, Latin BERT, is proposed by (Bamman and Burns, 2020). They pre-trained BERT model on Latin texts from Perseus, PROIEL and Index Thomisticus Treebank, targeting restoration and several other downstream tasks. (Sommerschield et al., 2023) offer a review on published research using machine learning for the study of ancient texts. They also classify the studies of ancient texts into tasks: digitisation, restoration, attribution, linguistic analysis, textual criticism, translation and decipherment. Finally, a similar review task takes place in the article (Braović et al., 2024), but focusing on the computational techniques related to the Bronze Age Aegean and Cypriot scripts, namely the Archanes script and the Archanes formula, Cretan hieroglyphic (including the Malia Altar Stone and Arkalochori Axe), Phaistos Disk, Linear A, Linear B, Cypro-Minoan and Cypriot scripts.

The work in (Papavassileiou et al., 2023) is similar to ours and is the only one that we are aware of that does infilling for the Linear B tablets. However, that work is limited to series D tablets and considers both phonogrpahic and non-phonographic symbols. Furthermore, like all aforementioned methods, it does not investigate transfer learning.

3 The Mycenaean dataset

Here we present the modifications we made to the Mycenaean dataset of series D, that was initially created as described in (Papavassiliou et al., 2020) and (Papavassileiou et al., 2023). We also present the way to create the new dataset of series A&B.

The Linear B script uses two basic symbol systems, one phonetic (phonographic component) and one logographic (non-phonographic component). The symbols of the phonetic system are called syllabograms-syllables. The phonetic system is usually represented transcribed, i.e., the syllable is rendered in letters, and in most cases by a combination of consonant and vowel. The system of the phonetic symbols, includes at least 87 different syllables. For the symbols of the logographic system, the term 'ideograms' or 'logograms' is used, sometimes modified by ligatured signs or 'adjuncts' (mostly acrophonic abbreviations) (Petrakis, 2017). The ideograms are 143. For their representation a transcription is used, based on the abbreviation of the Latin name of the represented object or being, e.g., VIR 'man', MUL(ier) 'woman'. Additionally

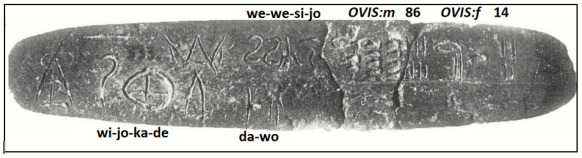

Figure 1: The image of the Mycenaean Linear B tablet KN Dd 1155 + 5378 + 5688 (the copyright of the images belongs to the Hellenic Ministry of Culture and Sports - "Hellenic Organization of Cultural Resources Development"). Translation of the Mycenaean tablet "Wiokados (shepherd's name): at Dawos (place name) belonging to Werwesios (collector's name), rams 86, ewes 14"

there are numbers that follow the decimal system and measurements units of weight and capacity (Ruiperez and Melena, 1996; Duhoux, 2014).

The assumption for the creation of both datasets was based on the clear separation of the signs of the Linear B writing system into phonograms and non-phonograms. The phonographic part is made up of the syllables, while the non-phonographic component includes the ideograms sometimes modified by ligatured signs or 'adjuncts' (mostly acrophonic abbreviations), as well as signs for measurement units, numbers, and other marks, such as signgroup dividers or 'check-marks' (Petrakis, 2017). Therefore, we decided to exclude the ideograms.

3.1 The dataset of series D

The series D of Knossos, which is the largest classification, comprises the accounts of sheep herds in around 1100 tablets. Most of them were probably written by the same scribe and have a similar structure. From those tablets 513 complete sequences were extracted, without missing syllables. From the augmentation rules (similar to (Papavassileiou et al., 2023)) were obtained 2052 sequences (725 augmented samples and 1327 duplicated samples). So, 2565 constitute the training set of the model. The remaining sequences were more or less damaged, making about 145 sentences. The new Vocabulary was defined by [77 syllables, space].

In essence, the modification we made in the dataset of series D, (Papavassileiou et al., 2023), concerns the removal of ideograms from the sequences, along with the numeric signs and the measurement units of weight and capacity. Thus, only the signs which occur in groups, i.e., the words, make up the new corpus. An example of such a sequence is shown in Table 1 derived from Figure 1.

Mycenaean sequence with ideograms
wi-jo-ka-de da-wo we-we-si-jo $OVIS^m$ $OVIS^f$
Mycenaean sequence without ideograms
wi-jo-ka-de da-wo we-we-si-jo

Table 1: Mycenaean sequence extracted from the Mycenaean Linear B tablet KN Db 1155 + 5378 + 5688 (Fig. 1) including ideograms (up), excluding ideograms (down).

3.2 The dataset of series A&B

We chose to include in the dataset the documents of series A&B found in Knossos (site of origin), to facilitate transfer learning from series D tablets, which originate from Knossos.

The tablets of series A&B write on staff lists/ staff statuses/ personnel situations; more specifically they include work groups. The introductory words, describing these groups, can be either a Cretan place-name, or a man's name (sometimes in genitive), or a feminine ethnic adjective from a place-name (ethnic-name), or an occupational name (trade-name) or some combination of these.

Some specific rules applied in this series are:

1. The tablets that contain complex (compound) sentences, are converted into simple ones. E.g., tablet KN Ai 63, Figure 2, writes "pe-se-re-jo e-e-si MUL 1 ko-wo 1 ko-wa 1" translated as "To Psellos belong one woman, one girl and one boy" (family or chattel slavery record). This tablet provides 3 sequences for our dataset: "pe-se-ro e-e-si", "pe-se-ro e-e-si ko-wa" and "pe-se-ro e-e-si ko-wo".

2. The second rule has to do with abbreviations. Most of the time the syllables are placed one after the other to form recognizable words. But, there are also cases where the syllables are used individually. When this happens, the syllable functions either as a ligature with an ideogram, or as an ideogram adjunct, or as an abbreviation of a word. There are numerous such annotations in series A&B that refer to the third case, abbreviations. In cases where we know the full form of abbreviated words, then the abbreviations are replaced by the full words. E.g., the tablet KN Ak 627, Figure 3, writes "da-*22-to a-no-zo-jo TA 1 DA 1 MUL 9 pe di 2 ko-wa me-zo-e 7 ko-wa me-wi-jo-e 10 ko-wo me-zo-e 2 ko-wo me-wi-jo-e 10". Here, the abbreviations 'pe' and 'di' appear.

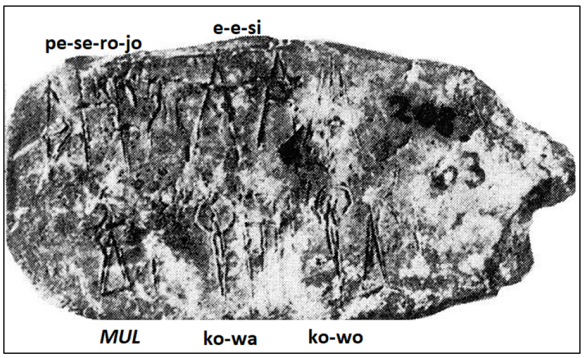

Figure 2: The image of the Mycenaean Linear B tablet KN Ai 63(image copyright belongs to the Hellenic Ministry of Culture and Sports - "Hellenic Organization of Cultural Resources Development"). Translation of the Mycenaean tablet *To Psellos (name of a person) belong one woman, one girl and one boy*.

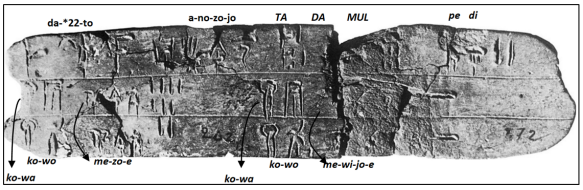

Figure 3: The image of the Mycenaean Linear B tablet KN Ak 627 + 7025 + fr (image copyright belongs to the Hellenic Ministry of Culture and Sports - "Hellenic Organization of Cultural Resources Development"). Translation of the Mycenaean tablet *In area da-*22-to the work group belonging to a-no-zo consists of one leader, one damar, nine women, of which two come from the censuses of the previous year in a period of apprenticeship, 7 older girls, 10 younger girls, 2 older boys and 10 younger boys*.

> These are the shortened forms of the words 'pe-ru-si-nu-wo' (last year's) and 'di-da-ka-re' (during apprenticeship/under instruction) (Ventris and Chadwick, 2015), which will appear in their full form in the dataset.

We gathered all the sequences that emerged from the tablets of series A&B, without missing symbols, 426 in number. The sequences resulting from the damaged tablets of this category were around 159. We defined a Vocabulary of [75 syllables, space].

The augmentation used in series D cannot be applied to series A&B tablets due to the fact that in A&B we mostly encounter lists of human names.

4 Supervised learning for infilling Mycenaean tablets

In the first place, we employ a symbol-level Bidirectional Recurrent Neural Network (BRNN), (as has been employed in (Papavassileiou et al., 2023)) to fill in the gaps in the Mycenaean tablets of series D and A&B. The goal in this case is to check if the predictions of the model improve by removing the ideograms from the Mycenaean sequences.

4.1 Modeling series D

A variation of Leave One Out Cross Validation (LOOCV) procedure was followed, the *Leave-"one and its derivatives"-Out Cross Validation*, in order to evaluate the BRNN model on unseen data, considering the scarcity of data. The derivatives are the samples/sequences resulting from applying an augmentation step to the real sample/sequence that is currently left out for testing. So, the samples resulting from augmentation of the current test sample sequences were excluded from the respective training set to avoid contamination of the test set by including sample sequences of the same origin (through augmentation). Furthermore, only the original sequences were used for testing (not the augmented or the duplicated ones), to make results comparable to those where no augmented data were used. Thus, the model is trained 513 times and the final performance is based on all these runs.

We implemented a function that performs one step of stochastic gradient descent with gradient clipping, $ClippingValue = 0.5$. We applied the greedy heuristic approach to search for the best hyperparameters, ending up with: 110.000 iterations ($epochs = 43$), number of neurons in hidden layer $N_{hl} = 57$, and learning rate $l_r = 0.01$. As an output activation function was set the softmax function and for the hidden layer the hyperbolic tangent (Tanh) function was chosen. For the initialization of weight matrices and bias vectors, we ended up in *"Glorot/Xavier"* as the most suitable for use. Given those choices we came to the results for the BRNN shown in Table 2, which demonstrate an improvement when compared to the baseline.

We used the trained model to infill gaps for which experts made educated guesses on the missing parts (Chadwick et al., 1987), mainly based on the visual cues, since some small parts of the syllables remain visible. The experts didn't use the sequences' structure unlike our method. Eight (8) of our TOP-5 predictions agree with the literature recommendations. This number shows an improvement of 2 units compared to the corresponding training of the BRNN model on the dataset including ideograms (Papavassileiou et al., 2023). See the Appendix A for more details.

SERIES D - with ideograms				
TOP-1	**TOP-5**	**TOP-10**	**TOP-15**	**TOP-20**
48.34	65.30	71.93	74.85	78.17
SERIES D - without ideograms				
TOP-1	**TOP-5**	**TOP-10**	**TOP-15**	**TOP-20**
48.56	65.50	72.12	76.02	80.12

Table 2: Estimated scores (percentages) of finding the correct missing symbol among the top-k most likely symbols (k=1,5,10,15,20) according to the probabilities estimated by the BRNN model. Up, the training dataset includes the ideograms. Down, the training dataset does not include the ideograms.

4.2 Modeling series A&B

At this point we evaluate the performance of the model learnt from A&B series. In a similar fashion to D series, we conducted two experiments, one with synthetic gaps and one with real ones.

4.2.1 Infilling synthetic sequences

To estimate the performance of the model on unseen data, the Leave-One-Out Cross-Validation (LOOCV) is used, since the data are scarce.

We randomly removed syllables from the 426 sequences in order to test the prediction capability of the BRNN model. The creation of the synthetic gaps follows the distribution of the real gaps appearing in the damaged tablets of this category. 67% of the real gaps occur at the beginning of the sequence, 17% somewhere in the middle of the sequence and 26% in the end of the sequence. We created a similar distribution for synthetic gaps.

The model is trained 426 times and the final performance estimate is based on all these runs. We used the Cross-Entropy loss along with the stochastic gradient descent optimizer with gradient clipping value of 0.5. The network is trained for 55.000 iterations ($epochs = 129$) - $batch-size = 1$, with 58 neurons in the single hidden layer and learning rate, $lr = 0.01$. For the hidden layer of the neural network was used the hyperbolic tangent (Tanh) activation function and as an output activation function was chosen the softmax function since it is a d-way classification problem. Finally, *"Glorot/Xavier"* was chosen for the initialization of weight matrices and bias vectors.

Given those choices we came up with the results in Table 3. The model's prediction rates on series A&B are lower than those of its counterpart (Table 2 (down)) on series D. This is mainly due to the fact that the tablets of series A&B offer fewer sequences and we could not formulate augmentation rules.

SERIES A&B				
TOP-1	**TOP-5**	**TOP-10**	**TOP-15**	**TOP-20**
30.28%	50.23%	57.75%	61.97%	66.20%

Table 3: Estimated scores (percentages) of finding the correct missing symbol among the top-k most likely symbols (k=1,5,10,15,20) according to the probabilities estimated by the BRNN model. The training dataset does not include the ideograms.

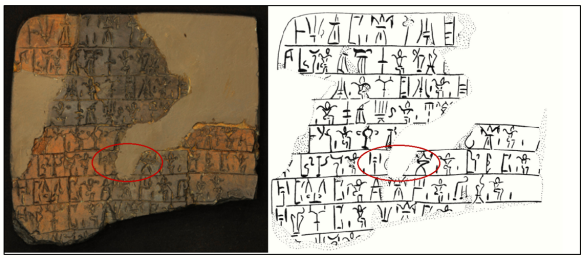

Figure 4: The image of the damaged Mycenaean tablet KN Bk 799 + 8306. (© Hellenic Ministry of Culture) (left) and its drawing (right). It lists men's names. Thirteen of them are complete. Around 6 names remain unknown.

4.2.2 Infilling real sequences

In this experiment, the model is applied in some real cases. In order to predict the missing syllables from the 159 sequences obtained from the damaged tablets of series A&B, we used the BRNN model from the previous experiment (A) which was trained on the 426 complete sequences.

In this category, experts have offered their opinion on a number of cases, 29 in all, as to what the missing syllable might be, based on the visual similarity that the remnant might have with the Mycenaean syllables. For example, experts suggest that the residue on the tablet in Figure 4 is equally likely to match the syllables 'ka' and 'qe', thus completing the man's name 'a-ka-de' or 'a-qe-de'. More such results are presented in the Appendix A.

Of the model's TOP-10 predictions, slightly more than half (15) match the experts' estimates. Of these correct predictions, always in agreement with the visual assessments of the experts, 9 are in the top 5 predictions (TOP-5), Table 5. This gives an indication that the model can learn effectively from the data. More details are given in the Tables of the Appendix A.

One way to increase the data of series A&B is to include the tablets of this series from other sites, namely Pylos, Thebes, Mycenae etc. However, that process has to be done with great caution so as not to contaminate the content of the dataset.

5 Transfer learning for infilling series A&B

In the following we investigate different transfer learning approaches for enhancing the model for series A&B (target) using the model learnt from series D (source). Series D data involves more tablets and is syntactically richer while the data in series A&B is scarce. However, for the NLP standards both series are considered very scarce.

We have essentially experimented with 2 simple TL techniques:

(A) Use the parameters of the model trained on D, see Section 4.1, as initialization for the training of the A&B series model (CASE A).

(B) Freeze the D model while training a second hidden layer using Tanh activation, finally fed to a softmax output layer (CASE B).

5.1 Case A: Weights initialization from D

We use the pre-trained D model only to initialize the A&B model.

By replacing, in the weight initialization procedure, the *"Glorot/Xavier"* method with the optimal parameters extracted from the training of another neural network model, we seek to examine whether there will be an improvement in the results.

The first goal is to use the optimal parameters extracted from the pre-trained model in series D corpus, as weight initialization for the training of series/corpus A&B.

In each LOOCV evaluation/iteration, the optimal parameters from the training of corpus D are used as weight initialization. The best results from the training of series D were collected with hyperparameters; learning rate = 0.01, mini batch size = 1, epochs = 43 (110.000 iterations), one hidden layer with 57 neurons/units.

We experimented with the number of iterations as hyperparameter and the results of Table 4 (CASE A) were achieved with 12.000 iterations ($epochs \approx 28$).

5.2 Case B: Add and train a second hidden layer

In this case we connected the D model with an additional neural network layer of 55 neurons (emerged after many tests), in order to train the A&B corpus, and to experiment with the following technique: Freeze the trained model in D and train the attached layer (CASE B).

CASE A				
TOP-1	**TOP-5**	**TOP-10**	**TOP-15**	**TOP-20**
25.65%	43.66%	54.23%	62.91%	67.37%
CASE B				
TOP-1	**TOP-5**	**TOP-10**	**TOP-15**	**TOP-20**
24.71	42.96	50.70	57.28	63.15

Table 4: Estimated scores (percentages) of finding the correct missing symbol among the top-k most likely symbols (k=1,5,10,15,20) according to the probabilities estimated by the TL models, CASE A and CASE B

Here we kept the weights of the pre-trained layer frozen while we trained only the attached neural network layer. The initial layer has 57 neurons, $N_{hl_1} = 57$, since it corresponds to the pre-trained model in series D, and the second layer has 55 neurons, $N_{hl_2} = 55$. In each iteration, $iters = 6.000$ ($epochs \approx 14$), the weights of the initial layer remain frozen, while the weights of the higher/second layer were readjusted/updated. Thus, we ended up with a bidirectional recurrent neural network with two hidden layers, which on the second layer performs one step of stochastic gradient descent with gradient clipping, $ClippingValue = 0.8$, and learning rate, $lr = 0.01$. Table 4 (CASE B) illustrates the results of this architecture.

5.3 Assessment

Comparing the experiments of TL with that of training on series A&B from scratch, we observe the following:

- There is no overall improvement in synthetic gap infilling in comparison to learning from scratch as displayed in Table 3 and Table 4.
- The results of the trained model of cases A and B on the 29 real cases of the series A&B are presented in Table 5. The TL again does not seem to outperform the model trained from scratch in A&B series. However, the model behaves better and actually gives solutions in some complex cases, in cases where only one syllable has survived from the incomplete word (e.g., Mycenaean tablets KN Ak 7022 [+] 7024 and KN Ai 7745), while the model from supervised learning does not. These are described in detail in Appendix A.

Surely further investigation is needed on TL methods. The relative success for the real gap infilling task let us assume that if the data of series D (pre-trained model) increases then we will probably get better prediction rates.

Series A&B	BRNN		TL CASE A		TL CASE B	
29 suggestions	**TOP5:** 9	**TOP10:** 15	**TOP5:** 9	**TOP10:** 12	**TOP5:** 8	**TOP10:** 14

Table 5: Number of predictions (TOP-5 and TOP-10) in agreement with the visual assessments of the experts in the 29 instances for the three cases concerning the training of category A&B.

6 Conclusions and future work

Our model exploits a character-level Bidirectional Recurrent Neural Network and two Transfer Learning approaches in order to capture the statistical structure of the Mycenaean documents. Our methodology is expected to assist the experts recover the missing parts by offering alternatives along with their probability, which are complementary to the visual channel. The key takeaways are described in the following.

Training the BRNN model on the different series D datasets, by excluding the ideograms, we experienced a small improvement over the with-ideograms dataset.

The training of a similar BRNN model in series A&B from Knossos gives reasonable results. The prediction rates are reasonably lower, since the dataset includes significantly fewer tablets and consequently offers fewer sequences; these sequences are much shorter, most of them a single word, compared to those of series D.

We explored the potential of transfer learning techniques in a small dataset, with mixed results. Although the overall performance is not better than training from scratch, the TL should not be rejected because it exhibits some complementarity with supervised learning. Further investigation is needed, potentially with more data series.

The research can be extended to incorporate more series (apart from series D and A&B there are about 12 more series to investigate), including newly discovered or previously unexplored Mycenaean tablets. Increasing the size and diversity of the dataset can contribute to the robustness and generalization of the models, enabling them to handle a broader range of linguistic variations and complexities. Furthermore, we can incorporate Mycenaean tablets from other sites, not only from Knossos, for example from Pylos, Thebes, etc. Such an attempt will not only increase the dataset but will also contribute to enhancing the diversity of the data.

The incorporation of the visual modality is another aspect that we have not investigated so far, but should do in our next steps.

References

Yannis Assael, Thea Sommerschield, and Jonathan Prag. 2019. Restoring ancient text using deep learning: a case study on Greek epigraphy. In *Empirical Methods in Natural Language Processing*, pages 6369–6376.

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Maria Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603:280–283.

David Bamman and Patrick J. Burns. 2020. Latin BERT: A contextual language model for classical philology. *CoRR*, abs/2009.10053.

A. Bernabé and E.R. Luján. 2020. *Introducción al griego micénico. Gramática, selección de textps y glosario*. Monografías de Filología Griega. Prensas de la Universidad de Zaragoza.

Maja Braović, Damir Krstinić, Maja Štula, and Antonia Ivanda. 2024. A systematic review of computational approaches to deciphering bronze age aegean and cypriot scripts. *Computational Linguistics*, pages 1–54.

J. Chadwick. 1990. *The Decipherment of Linear B*. Canto. Cambridge University Press.

J. Chadwick, L. Godart, J. T. Killen, J. P. Olivier, A. Sacconi, and I. A. Sakellarakis. 1987. *Corpus of Mycenaean Inscriptions from Knossos: Volumes 1-4*. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mario. Doria. 1965. *Avviamento allo studio del Miceneo : struttura, problemi e testi / Mario Doria.* Incunabula Graeca ; v.8. Edizioni dell'Ateneo, Roma.

Y. Duhoux. 2014. *A companion to linear B : Mycenaean Greek texts and their world*. Number v. 3 in Series: Bibliothèque des Cahiers de l'Institut de Linguistique de Louvain (BCILL), 133. Peeters.

Y. Duhoux and A.M. Davies. 2008. *A Companion to Linear B: Mycenaean Greek Texts and Their World*. Bibliothèque des Cahiers de l'Institut de Linguistique de Louvain. Peeters.

Ethan Fetaya, Yonatan Lifshitz, Elad Aaron, and Shai Gordin. 2020. Restoration of fragmentary babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences*, 117(37):22743–22751.

Margalit Fox. 2013. *Riddle of the labyrinth: the quest to crack an ancient code and the uncovering of a lost civilisation*. Profile Books.

M.D. Freo and M. Perna. 2019. *Manuale di epigrafia micenea: introduzione allo studio dei testi in lineare B*. Libreriauniversitaria.it edizioni.

Kyeongpil Kang, Kyohoon Jin, Soyoung Yang, Soojin Jang, Jaegul Choo, and Youngbin Kim. 2021. Restoring and mining the records of the Joseon dynasty via neural language modeling and machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4031–4042, Online. Association for Computational Linguistics.

J. T. Killen. 1964. The interpretation of mycenaean greek texts by l. r. palmer. oxford: the clarendon press, 1963. 501 pp., frontispiece, 4 figs. 70s. *Antiquity*, 38(150):148–150.

Alice Elizabeth Kober. 1945. Evidence of inflection in the "chariot" tablets from knossos. *American Journal of Archaeology*, 49:143 – 151.

Alice Elizabeth Kober. 1946. Inflection in linear class b: 1-declension. *American Journal of Archaeology*, 50:268 – 276.

Alice Elizabeth Kober. 1948. The minoan scripts: Fact and theory. *American Journal of Archaeology*, 52:82 – 103.

Koren Lazar, Benny Saret, Asaf Yehudai, Wayne Horowitz, Nathan Wasserman, and Gabriel Stanovsky. 2021. Filling the gaps in Ancient Akkadian texts: A masked language modelling approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4682–4691, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Torsten Meissner. 2001. F. m. j. waanders: Studies in local case relations in mycenaean greek. pp. vii 134. amsterdam: J. c. gieben, 1997. paper, hfl. 65. isbn: 90-5063-107-x. *The Classical Review*, 51(1):179–180.

The Packard Humanities Institute. 2005. PHI Greek inscriptions. https://inscriptions.packhum.org/. Online accessed 2021.

Katerina Papavassileiou, Dimitrios I. Kosmopoulos, and Gareth Owens. 2023. A generative model for the mycenaean linear b script and its application in infilling text from ancient tablets. *J. Comput. Cult. Herit.*, 16(3).

Katerina Papavassiliou, Gareth Owens, and Dimitrios Kosmopoulos. 2020. A dataset of Mycenaean Linear B sequences. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2552–2561, Marseille, France. European Language Resources Association.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Vassilis Petrakis. 2017. Figures of speech? observations on the non-phonographic component in the linear b writing system. In *Aegean Scripts: Proceedings of the 14th International Colloquium on Mycenaean Studies*, v. 105, pt. 1, pages 127–167. Edizioni Consiglio nazionale delle ricerche, Istituto di studi sul Mediterraneo antico.

Maurice Pope. 2008. The decipherment of linear b. In Anna Morpurgo Duhoux, Yves; Davies, editor, *A Companion to Linear B: Mycenaean Texts and their World*, volume 1, pages 3–11. Louvain-la-Neuve, Belgium: Peeters.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rajesh P. N. Rao, Nisha Yadav, Mayank N. Vahia, Hrishikesh Joglekar, R. Adhikari, and Iravatham Mahadevan. 2009. A markov model of the indus script. *Proceedings of the National Academy of Sciences*, 106(33):13685–13690.

A. Robinson and S. Eisenman. 2002. *The Man who Deciphered Linear B: The Story of Michael Ventris*. World of Art Series. Thames & Hudson.

Henriette Roued-Cunliffe. 2010. Towards a decision support system for reading ancient documents. *Literary and Linguistic Computing*, 25(4):365–379.

C.J. Ruijgh. 1977. E. L. Bennett J.P. Olivier, the Pylos tablets transcribed, i : Texts and notes, ii : Hands, concordances, indices (incunabula graeca, 51 et 59). Rome, edizioni dell'Ateneo, 1973 et 1976. 287 et 146 p. pr. l. 10.600 et 14.000. *Mnemosyne*, 30(3):296 – 298.

M. S. Ruiperez and J. L. Melena. 1996. *The Mycenaean Greeks*. Athens: Kardamitsa.

Tianxiao Shen, Victor Quach, Regina Barzilay, and Tommi Jaakkola. 2020. Blank language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5186–5198, Online. Association for Computational Linguistics.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine Learning for Ancient Languages: A Survey. *Computational Linguistics*, pages 1–44.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27, pages 3104–3112. Curran Associates, Inc.

M. Ventris and J. Chadwick. 1953. *Evidence for Greek Dialect in the Mycenaean Archives*. Journal of Hellenic Studies. Journal of Hellenic Studies.

M. Ventris and J. Chadwick. 1956. *Evidence for a Greek Dialect in the Mycenaean Archives*. Council of the Society for the Promotion of Hellenic Studies.

M. Ventris and J. Chadwick. 2015. *Documents in Mycenaean Greek*. Cambridge University Press.

M. Ventris, A. Sacconi, and J. Chadwick. 1988. *Work Notes on Minoan Language Research and Other Unedited Papers*. Incunabula Graeca. Edizioni dell'Ateneo.

Krzysztof Wolk and Krzysztof Marasek. 2014. Polish-English speech statistical machine translation systems for the IWSLT 2014. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 143–149, Lake Tahoe, California.

Nisha Yadav, Hrishikesh Joglekar, Rajesh P. N. Rao, Mayank N. Vahia, Ronojoy Adhikari, and Iravatham Mahadevan. 2010. Statistical analysis of the indus script using n-grams. *PLoS ONE*, 5(3):e9506.

A Appendix

Here we present the damaged Mycenaean tablets, along with the experts' guesses based on the visual cues of some small parts of the missing syllables (Chadwick et al., 1987). We compare our models (Bidirectional Recurrent Neural Network and Transfer Learning) with those estimations.

The bibliographic comments on the missing symbol, are shown in the third column, and the results of the model in the fourth one. The symbol *'bl'* corresponds to the 'space' symbol.

In Table 1 are presented the D series BRNN predictions for the 15 cases for which we have the opinions of the experts. Eight of our TOP-5 predictions agree with the literature recommendations. In the following we try to give some possible explanations for the remaining 7 model's predictions, since there is no hard evidence available:

- On the tablet Dl 933 we find a syllable, '*83' (its phonetic value has not been determined with certainty), which is quite rare and is observed only twice in the training dataset. As a result, the model is probably not sufficiently trained in such a context.
- The TOP-5 BRNN predictions are not consistent with the remnant in the KN Dv 1213 and KN Dv 5236 + 532. The model prioritizes other syllables, probably due to the short sequence length, which conveys rather poor context information. Inclusion of visual evidence in our model in the future could handle such issues.
- The tablet Da 1341 is a difficult case, since the visual evidence is very weak and even the experts note that their degree of certainty is low.
- For the tablet KN Db 5310 + 6062 + 837 we notice that the predictions of our model, and in fact the 'ka' and 'ra', are not completely irrelevant/unrelated to the remnant.
- The tablet Db 5359 + 5565 +7214 is another difficult case as acknowledged by experts and their degree of certainty is also low.
- The remnant in tablet KN Do 7740 largely matches the syllable 'ke', suggested by the bibliography. The model in this case probably failed to predict part of a human name; human names are typically unique.

As regards Tables 2, 3 and 4 related to the experts' suggestions in the 29 real cases of series A&B, we observe the following:

- There is a convergence of the models after Transfer Learning with the estimations of the experts for real gaps. If we observe the predictions of the BRNN model on the damaged tablets of series A&B, Table 2, in relation to those after applying the Transfer Learning technique, Tables 3 and 4, we will see that when the predictions of all three models agree with the bibliographic annotation, tablets KN Bk 806 + 6053 + frr (eighth row) and KN As 1516 (thirteenth row), then the predictions of the TL models rank higher (among the TOP-5 predictions).
- Another important observation, concerning the TL method, results from the tablet KN Ak 7022 [+] 7024. This tablet offers 6 sequences (twenty-first - twenty-sixth row), Tables 2, 3 and 4. The bibliography for the incomplete word, "*-ki", suggests the syllable "do". This syllable begins to appear in the BRNN predictions of the Table 4, which concerns TL method (CASE B). It is quite difficult to predict the rest of a word when there is only one syllable left. These cases are more likely to be approached with TL techniques rather than training from scratch.
- It is no coincidence that results for tablet KN Ai 7745 (twenty-ninth row), in agreement with the experts' opinion, we only have with TL method and indeed in CASE B (Table 4) it is the first choice.
- Something similar happens with tablets KN Bk 5134 (fifteenth row) and KN As 5932 [+] 8342 (twentieth row). Only the TL method manages to display the desired syllable in the TOP-10 predictions, Table 4 for the KN Bk 5134 and Table 3 for the KN As 5932 [+] 8342, with the difference that they are not in the TOP-5.

Due to these reasons we believe that TL methods need further investigation.

Damaged Tablets	**Sequences**	**Bibliographic annotation**	**BRNN TOP-5**
KN Dq 447	*-ta-wo da-mi-ni-jo	*possibly 'ka' or 'qe'*	***'ka'***, *'si', 'ku', 'qo', 'ti'*
KN Dl 933 + 968 + 975	*-*83-re-to si-ja-du-we po-ti-ni-ja-we-jo	*perhaps 'ko'*	*'no', 'a3', 'ra', 'wi', 'do'*
KN Dp 1061	*-sa pa-i-ti-ja	*probably 'to'*	***'to'***, *'sa', 'te', 'po', 'qa'*
KN Dv 1213	*-so u-ta-no	*'to', 'jo' possible*	*'pu', 'ko', 'nwa', 'i', 'wi'*
KN Da 1341 [+] 1454 + 8777	*-no-qa-ta pa-i-to da-mi-ni-jo	*'po' possible, but difficult*	*'a', 'wi', 'jo', 'ka', 'ti'*
KN Db 1344 + 6017 + 7268 + 7950 + 8235	*-tu-to pa-i-to we-we-si-jo-jo	*perhaps 'ti'*	*'bl', 'ku', 'ra',* ***'ti'***, *'pa'*
KN Da 1401 + 7998 + fr.	*-wi da-*22-to	*perhaps 'na'*	*'ro', 'ri',* ***'na'***, *'nu', 'ra'*
KN Df 5198 + 5238 + 5269	wi-na-jo ra-* ki-ri-jo-te	*traces favour 'ja'*	*'to',* ***'ja'***, *'su', 'nwa', 'pe'*
KN Dv 5236 + 5329	*-jo ra-to	*perhaps 'qa' or 'wo'*	*'te', 'ku', 'ri', 'wa', 'ro'*
KN Dv 5278 + 5338 + 8557	*-ma-we qa-mo	*'ko' not impossible*	*'ja', 'i',* ***'ko'***, *'sa', 'no'*
KN Db 5310 + 6062 + 8375	e-*-jo ku-ta-to	*perhaps 'qa' or 'ri'*	*ka', 'ra', 'wa', 'te', 'se'*
KN Db 5359 + 5565 + 7214	*-ma-na-so ra-su-to u-ta-jo	*'pi' not impossible*	*'bl', 'ta', 'wa', 'ru', 'di'*
KN Dv 5690	du-ni-*	*'jo' possible*	***'ja'***, *'to', 'po', 'ki', 'mi'*
KN Dc 7161 + 7179 + 8365 + fr.	*-to ku-ta-to u-ta-jo-jo	*possibly 'ke'*	*'ra', 'ta',* ***'ke'***, *'ro', 'ko'*
KN Do 7740	*-ta ka-to-ro se-to-i-ja	*'ke' or 'de'*	*'u', 'ku', 'wa', 'si', 'qo'*

Table 1: Bibliographic annotations (Chadwick et al., 1987) comparing to BRNN predictions in the 15 sequences appear in the real cases of series D.

Damaged Tablets	Sequences	Bibliographic annotation	BRNN TOP-10
KN B 164 + 5666 + 7136 + 7544 + 8120 + frr.	o-da-*	*'ke' or 'je'*	*'wa', 'wi', 'ra', '*22', 'mo', 'ro', 'wo', 'mi', 'ma', 'zo'*
KN As 605 + 5869 + 5911 + 5931 + frr.	*-no pe-ro-qe	*'me' or 'ro'*	*'wo', 'tu', 'jo', 'ne', 'wi', 'ta',* ***'ro'****, 're', 'no', 'qa'*
KN Bk 799 + 8306 (Fig. 4)	a-*-de	*'ka' or 'qe'*	*'to', 'me', 'ke', 'pe', 'ta', 'mi', 'di', 'pi', 'ko', 'pa'*
KN Bk 802	ra-ti-*	*'jo'*	*'ja',* ***'jo'****, 'ri', 'pu', 'zo', 're', 'wa', 'nu', 'no', 'to'*
KN Bk 804	a-pa-re-*	*'u'*	*'jo', 'we',* ***'u'****, 'da', 'ta', 'ne', 'ti', 'te', 'wa', 'pa'*
KN Bk 806 + 6053 + frr.	ko-*-no	*'pi' or 'wi'*	*'wa', 'wo', 'no', 'to', 'so', 'ko', 'ma', 'qa', 'a', 'ta2'*
KN Bk 806 + 6053 + frr.	*-wo-ta	*'pi' or 'e'*	*'ne', 're', 'a', 'qi', 'wo', 'po', 'mo', 'ko', 'u', 'ri'*
KN Bk 806 + 6053 + frr.	ko-*-ka-ra-te-ne	*'wo'*	*'tu', 'we', 'bl', 'so', 'no', 'wa', 'da',* ***'wo'****, 'ma', 'u'*
KN B 809	*-sa-do-ro-jo	*'ke'*	***'ke'****, 'bl', 'to', 'pi', 'te', 'a', 'we', 'u', 'po', 'mi'*
KN As 1516	ko-no-si-ja ra-wa-ke-si-ja *-ki-wa-ta	*'a' or 'a3'*	***'a'****, 'e', 'pa', 'pi', 'mi', 'ni', 'du', 'wa',* ***'a3'****, 'do'*
KN As 1516	ko-no-si-ja ra-wa-ke-si-ja wa-du-*-to	*'ni' or 'sa'*	*'na', '*22', 'we', 'du', 'se',* ***'ni'****, 'de', 'ra2', 'za', 'nu'*
KN As 1516	*-ti-jo qa-si-re-wi-ja a-nu-to	*'ta' or 'ra'*	*'a', 'bl', 'do', 'i', 'po',* ***'ta'****, 'so', 'du', 'wa', 'o'*
KN As 1516	ku-*-ti-jo qa-si-re-wi-ja a-nu-to	*'ta' or 'ra'*	*'a', 'du', 'do', 'tu', 'i',* ***'ta'****, 'to', 'ku', 'ro2', 'au'*
KN As 1516	se-to-i-ja qa-si-re-wi-ja pi-*-jo	*'ri'*	*'si',* ***''ri'****, 'wi', 'ni', 'ra', 'u', 'mi', 'ro', 'mo', 'qi'*
KN Bk 5134	to-ke-*	*'u'*	*'qa', 're', 'zo', 'pu', 'ti', 'su', 'da', 'ka', 'no', 'ta'*
KN Bk 5172	*-wa-ta	*'ku'*	*'a', 'ki', 'ko', 're', 'i', 'po', 'wo', 'ne', 'ta', 'qi'*
KN Bk 5172	wi-do-*-wi	*'wo'*	*'e', 'ro', 'da',* ***'wo'****, 'ra', 'qe', 'ta', 'tu', 'to', 're'*
KN As 5609 + 6067	*-ke-u	*'i' or 'pa'*	*'e', 'ta', 'nu', 'ne', 'mi', 'wa', 'pi',* ***'pa'****, 'we', 'do'*
KN Am 5882 + 5902	*-so ka-ma-jo	*'to'*	***'to'****, 'no', 'tu', 'u', 'ni', 'ne', 'wo', 'wi', 'ko', 'ta2'*
KN As 5932 [+] 8342	a-*-we	*'ro'*	*'ke', 're', 'te', 'mi', 'pe', 'nu', 'pa', 'pi', 'ku', 'to'*
KN Ak 7022 [+] 7024	*-ki	*'do'*	*'to', 'jo', 'wo', 'bl', 'ni', 'ne', 'te', 'we', 'e', 'po'*
KN Ak 7022 [+] 7024	*-ki ko-wa me-zo-e	*'do'*	*'re', 'jo', 'e', 'wo', 'ni', 'ne', 'mi', 'to', 'qa', 'ri'*
KN Ak 7022 [+] 7024	*-ki ko-wa me-wi-jo-e	*'do'*	*'re', 'jo', 'e', 'wo', 'to', 'wi', 'ka', 'pu', 'mi', 'ni'*
KN Ak 7022 [+] 7024	*-ki ko-wo me-zo-e	*'do'*	*'re', 'to', 'wo', 'ni', 'tu', 'jo', 'mi', 'e', 'su', 'ri'*
KN Ak 7022 [+] 7024	*-ki ko-wo me-wi-jo-e	*'do'*	*'to', 're', 'jo', 'wo', 'ni', 'e', 'a', 'su', 'o', 'ri'*
KN Ak 7022 [+] 7024	*-ki do-e-ra	*'do'*	*'to', 're', 'qa', 'te', 'ke', 'ka', 'e', 'i', 'pi', 'qe'*
KN Bo 7043 + 7925	*-ra-so	*'ka' or 'qe' or 'we'*	***'ka'****, 'bl', 'ta',* ***'qe'****, 'si', 'pa', 'sa', 'a', 'wo', 're'*
KN Bg 7682	ri-si-*	*'jo'*	***'jo'****, 'ja', 'ra', 'nu', 'pu', 'mi', 'wi', 'no', 'ro', 'zo'*
KN Ai 7745	*-ja-to si-qa	*'ri'*	*'ti', 'i', 're', 'qe', 'si', 'ta', 'wi', 'ni', 'bl', 'me'*

Table 2: Bibliographic annotations (Chadwick et al., 1987) comparing to BRNN predictions in the 29 sequences appear in the real cases of series A&B.

Damaged Tablets	Sequences	Bibliographic annotation	BRNN TOP-10
KN B 164 + 5666 + 7136 + 7544 + 8120 + frr.	o-da-*	*'ke' or 'je'*	*'te', 'su', 'ka', 'ra', 'no', 'wi', 'wo', 'se', 'bl', 'so'*
KN As 605 + 5869 + 5911 + 5931 + frr.	*-no pe-ro-qe	*'me' or 'ro'*	*'a', 'bl', 'no', **'ro'**, 'i', 'ko', 'ka', 'ta2', 'qi', 'na'*
KN Bk 799 + 8306 (Fig. 4)	a-*-de	*'ka' or 'qe'*	*'ta', 'ra', 'to', 'pi', 'no', 're', 'ja', 'nu', 'wo', 'jo'*
KN Bk 802	ra-ti-*	*'jo'*	***'ja'**, 'bl', 'ke', 'ka', 'wa', 'ti', 'a', **'jo'**, 'qa', 'to'*
KN Bk 804	a-pa-re-*	*'u'*	*'so', 'ta', 'ka', 'te', 'se', 'si', 'wa', 'ti', 'we', 'po'*
KN Bk 806 + 6053 + frr.	ko-*-no	*'pi' or 'wi'*	*'a', '*56', 'ko', 'ta', 'bl', 'wo', 'mo', 'wa', 'no', 'ka'*
KN Bk 806 + 6053 + frr.	*-wo-ta	*'pi' or 'e'*	*'ro', 'to', 'da', 'jo', 'no', 'qo', 'nu', 'ko', 'du', 'di'*
KN Bk 806 + 6053 + frr.	ko-*-ka-ra-te-ne	*'wo'*	*'bl', 'no', 'a', **'wo'**, 'ja', 'wa', 'ta', 'si', 'i', 'e'*
KN B 809	*-sa-do-ro-jo	*'ke'*	*'bl', 'ta', 'ro', 'no', 'jo', 'ka', 'ja', 'ne', 'e', 'so'*
KN As 1516	ko-no-si-ja ra-wa-ke-si-ja *-ki-wa-ta	*'a' or 'a3'*	*'no', 'mi', 'ru', 'so', 'pe', 'da', 'si', '*56', 'to', 'do'*
KN As 1516	ko-no-si-ja ra-wa-ke-si-ja wa-du-*-to	*'ni' or 'sa'*	*'ri', 'se', '*22', 'qa', **'ni'**, 'su', 'ru', 'no', '*56', 'ka'*
KN As 1516	*-ti-jo qa-si-re-wi-ja a-nu-to	*'ta' or 'ra'*	*'i', 'do', 'se', 'ku', 'wi', 'po', 'na', 'tu', 'wa', 'a'*
KN As 1516	ku-*-ti-jo qa-si-re-wi-ja a-nu-to	*'ta' or 'ra'*	*'i', **'ta'**, 'se', 'do', 'qe', 'o', 'su', 'ra2', 'ko', 'nwa*
KN As 1516	se-to-i-ja qa-si-re-wi-ja pi-*-jo	*'ri'*	***'ri'**, 'ro', 'si', 're', 'sa', 'da', 'mi', 'ra', 'tu', 'ru'*
KN Bk 5134	to-ke-*	*'u'*	*'a', 'bl', 'ka', 'to', 'o', 'ja', 'e', 'ti', 'ru', 'ta'*
KN Bk 5172	*-wa-ta	*'ku'*	*'ro', 'bl', 'ki', 'di', 're', 'ko', 'so', 'u', 'a3', 'qo'*
KN Bk 5172	wi-do-*-wi	*'wo'*	*'ti', 'bl', 'so', **'wo'**, 'ro', 'si', 'su', 'sa', 'ne', 'o'*
KN As 5609 + 6067	*-ke-u	*'i' or 'pa'*	*'to', 'ro', 'te', 'nu', 'sa', 'e', 'ne', 'ke', 'de', 'zo'*
KN Am 5882 + 5902	*-so ka-ma-jo	*'to'*	***'to'**, 'po', 'na', 'ko', 'no', 'nu', 'du', 'pu', 'ro', 'e'*
KN As 5932 [+] 8342	a-*-we	*'ro'*	*'ta', 'ko', 're', 'te', 'ma', **'ro'**, 'ra', 'pi', 'nu', 'du'*
KN Ak 7022 [+] 7024	*-ki	*'do'*	*'si', 'jo', 'ro', 'to', 'no', 'bl', 'ta', 'ja', 'ne', 'so'*
KN Ak 7022 [+] 7024	*-ki ko-wa me-zo-e	*'do'*	*'to', 'te', 'jo', 'po', 'qo', 'e', 'na', 'si', 'no', 'me'*
KN Ak 7022 [+] 7024	*-ki ko-wa me-wi-jo-e	*'do'*	*'te', 'to', 'po', 'jo', 'qo', 'si', 'e', 'ro', 'we', 'no'*
KN Ak 7022 [+] 7024	*-ki ko-wo me-zo-e	*'do'*	*'to', 'te', 'po', 'jo', 'si', 'e', 'na', 'qo', 'no', 'ro'*
KN Ak 7022 [+] 7024	*-ki ko-wo me-wi-jo-e	*'do'*	*'te', 'po', 'jo', 'to', 'e', 'we', 'si', 'qo', 'na', 'no'*
KN Ak 7022 [+] 7024	*-ki do-e-ra	*'do'*	*'si', 'ta', 'no', 'ro', 'qa', 'di', 'wi', 'we', 'sa', 'i'*
KN Bo 7043 + 7925	*-ra-so	*'ka' or 'qe' or 'we'*	*'bl', 'ti', 'e', 'ta', 'ku', 'pa', 'zo', 'si', **'ka'**, 'wo'*
KN Bg 7682	ri-si-*	*'jo'*	***'ja'**, **'jo'**, 'no', 'ta', 'se', 'ni', 'mi', 're', 'ke', 'de'*
KN Ai 7745	*-ja-to si-qa	*'ri'*	*'ra', **'ri'**, 'ni', 'ti', 'na', 'i', 'mi', 'bl', 'ku', 're'*

Table 3: Bibliographic annotations (Chadwick et al., 1987) comparing to TL predictions in the 29 sequences appear in the real cases of series A&B (CASE A).

Damaged Tablets	**Sequences**	**Bibliographic annotation**	**BRNN TOP-10**
KN B 164 + 5666 + 7136 + 7544 + 8120 + frr.	o-da-*	*'ke' or 'je'*	*'ra', 'ko', 'wo', 'zo', 'su', '*22', 'to', '*56', 'sa', 'wa'*
KN As 605 + 5869 + 5911 + 5931 + frr.	*-no pe-ro-qe	*'me' or 'ro'*	*'a', 're', 'qa', 'to', 'da', 'wi', 'ki', 'su', 'ne', 'ti'*
KN Bk 799 + 8306 (Fig. 4)	a-*-de	*'ka' or 'qe'*	*'to', 'ta', 'no', 'po', 'te', 'nu', 'mi', 'tu', 'pe', 'a'*
KN Bk 802	ra-ti-*	*'jo'*	*'ri', 'to', 'e', 'so', 'we', 'i', 'ti', 'qa', 'ni', 'di'*
KN Bk 804	a-pa-re-*	*'u'*	*'ta', 'so', 'i', 'wa', 'po', 'jo', 'si', 'we', **'u'**, 'ka'*
KN Bk 806 + 6053 + frr.	ko-*-no	*'pi' or 'wi'*	*'a', 'wo', 'ko', 'to', 'ta', 'no', 'ma', 'ro', 'du', 'po'*
KN Bk 806 + 6053 + frr.	*-wo-ta	*'pi' or 'e'*	*'a', 'we', 'ko', 'po', 'ra', 'u', 'da', 'qi', 'bl', 'o'*
KN Bk 806 + 6053 + frr.	ko-*-ka-ra-te-ne	*'wo'*	*'a', **'wo'**, 'to', 'ta', 'e', 'tu', 'o', 'bl', 'ku', 'da'*
KN B 809	*-sa-do-ro-jo	*'ke'*	*'u', 'bl', 'to', 'a', 'qa', 'pi', 'te', 'di', 'we', 'ka'*
KN As 1516	ko-no-si-ja ra-wa-ke-si-ja *-ki-wa-ta	*'a' or 'a3'*	*'pu', **'a'**, 'ro', 'pe', 'si', 'wi', 'pi', 'se', 'di', 'su'*
KN As 1516	ko-no-si-ja ra-wa-ke-si-ja wa-du-*-to	*'ni' or 'sa'*	*'se', 'si', 'ri', 'ke', '*22', 'we', 'to', 'po', 'o', 'pe', 'mi'*
KN As 1516	*-ti-jo qa-si-re-wi-ja a-nu-to	*'ta' or 'ra'*	*'pe', 'pa', 'wa', 'i', 'do', 'ni', 'du', **'ta'**, 'ma', 'su'*
KN As 1516	ku-*-ti-jo qa-si-re-wi-ja a-nu-to	*'ta' or 'ra'*	*'ta', 'su', 'po', 'to', 'pa', 'du', 'pe', 'i', 'ko', 'ro2'*
KN As 1516	se-to-i-ja qa-si-re-wi-ja pi-*-jo	*'ri'*	*'ri', 'ra', 'ni', 'da', 'ti', 'si', 'ki', 'wa', 'mi', 'pi'*
KN Bk 5134	to-ke-*	*'u'*	*'ko', 'ja', 'so', 'qa', 'ku', 'da', 'to', **'u'**, 'a', 'ru'*
KN Bk 5172	*-wa-ta	*'ku'*	*'we', 'a', 'po', 'bl', 'si', 'o', 'ki', 'te', 'ti', 're'*
KN Bk 5172	wi-do-*-wi	*'wo'*	*'pe', 'e', 'sa', 'po', 'ra', 'do', 'te', 'zo', 'ti', 'pa'*
KN As 5609 + 6067	*-ke-u	*'i' or 'pa'*	*'pe', 'te', 'ri', 'ke', 'no', 'ro', 'se', 'zo', 'wo', 'do'*
KN Am 5882 + 5902	*-so ka-ma-jo	*'to'*	*'to', 'ta', 'ke', 'do', 'sa', 'ni', 'qa', 'te', 'ne', 'e'*
KN As 5932 [+] 8342	a-*-we	*'ro'*	*'ko', 'pe', 'te', 'po', 'nu', 'ke', 'a', 'da', 'me', 'no'*
KN Ak 7022 [+] 7024	*-ki	*'do'*	*'jo', 'pe', 'ni', 'wo', 'tu', 'te', 'no', 'to', 'ja', 'ta'*
KN Ak 7022 [+] 7024	*-ki ko-wa me-zo-e	*'do'*	*'ni', 'ka', 'wi', 'tu', 'mi', 'te', 'na', 'pe', 'to', 'ki'*
KN Ak 7022 [+] 7024	*-ki ko-wa me-wi-jo-e	*'do'*	*'ni', 'mi', 'te', 'tu', 'na', 'wi', 'ki', 're', 'e', 'ti'*
KN Ak 7022 [+] 7024	*-ki ko-wo me-zo-e	*'do'*	*'to', 'ka', 'ni', 'na', 'wi', 'tu', 'pe', 'te', **'do'**, 'ja'*
KN Ak 7022 [+] 7024	*-ki ko-wo me-wi-jo-e	*'do'*	*'ni', 'ka', 'to', 'na', 'wi', 'tu', 'pe', 'ki', 'se', **'do'***
KN Ak 7022 [+] 7024	*-ki do-e-ra	*'do'*	*'ka', 'ki', 'to', **'do'**, 'wi', 'su', 'mi', 'ti', 'tu', 'ni'*
KN Bo 7043 + 7925	*-ra-so	*'ka' or 'qe' or 'we'*	***'ka'**, 'si', 'zo', 'se', 'wo', **'qe'**, 'ta', 'a', 'nu', 'mo'*
KN Bg 7682	ri-si-*	*'jo'*	*'ja', 'wo', **'jo'**, 'ra', 'ni', 'to', 'ta', 'ti', 'de', 'ri'*
KN Ai 7745	*-ja-to si-qa	*'ri'*	***'ri'**, 'ro', 'ni', 'di', 'ti', 'mi', 'ke', 'pe', 'te', 're'*

Table 4: Bibliographic annotations (Chadwick et al., 1987) comparing to TL predictions in the 29 sequences appear in the real cases of series A&B (CASE B).