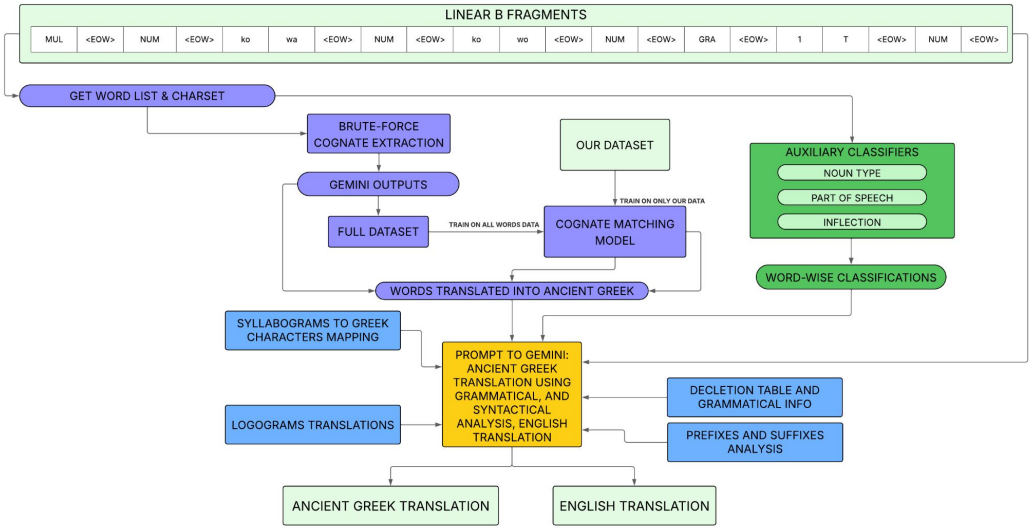
An AI Framework for Linear B Translation into Ancient Greek and English

SAPIENZA
UNIVERSITÀ DI ROMA

AI Framework for Linear B translation: overview

This presentation covers the following topics:

- What Linear scripts are and their data sources
- Cognate matching task between Linear B and Ancient Greek



- Auxiliary information gathering to perform translation
- A translation pipeline for Linear B automated translation into Ancient Greek and English

Aegean Linear Scripts: web scraping

- Cretan Hieroglyphic (2100 - 1700 B.C.)
- Linear A (1900 - 1450 B.C.)
- Linear B (1400 - 1200 B.C.)

→ https://sigla.phis.me/

774 full documents collected

https://liber.cnr.it/

5638 full documents collected

- Document level information
- Sequence level information
- Sign level information

Signs preprocessing and normalization

- Sign
- Function (LA)
- Sign number

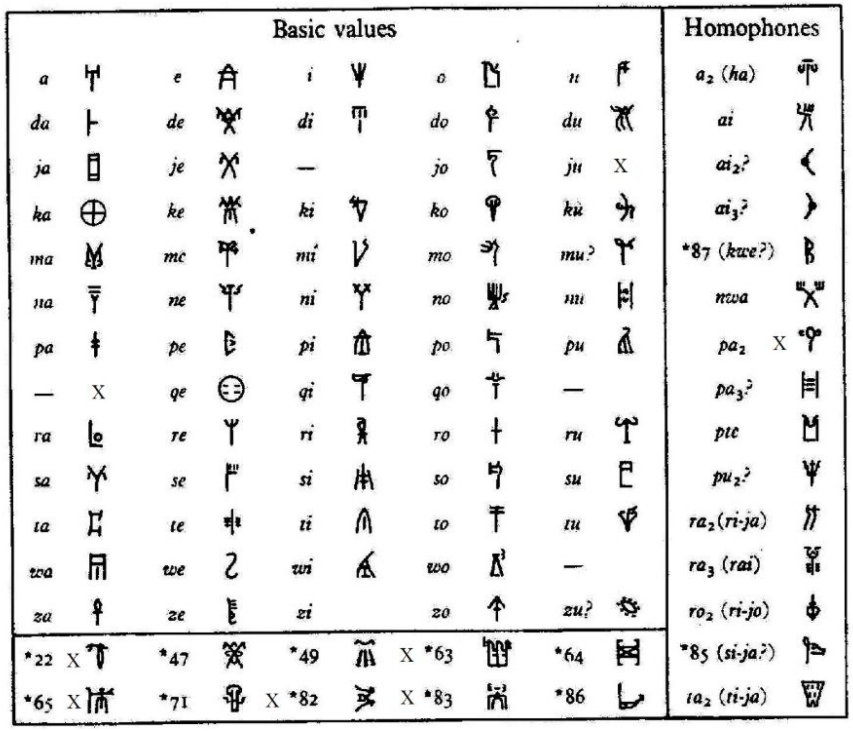
- Sequence
- Length
- Complete
- Findspot
- Sequence number

- Document name
- Link
- Chronology
- Findspot
- Size, support, motif (LA)
- Scribe, palmprint, museum inventory number (LB)

Linear B: the script

Linear B script is logosyllabic:

- logograms are symbols that stand for entire words
- syllabograms are symbols representing phonetic syllables



- Deciphered by Michael Ventris in 1952
- 70% of syllabograms shared with Linear A
- Used to write Ancient Greek

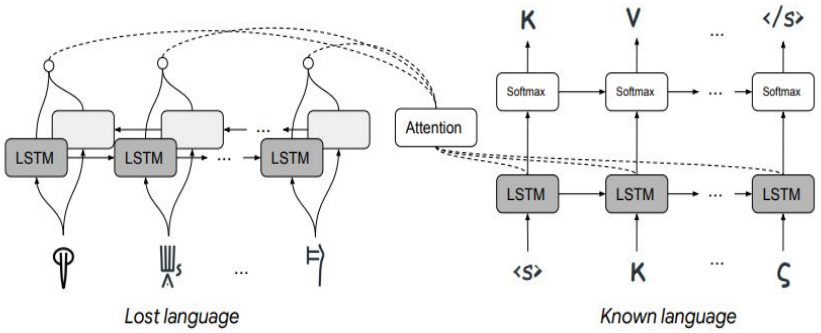
Cognate matching

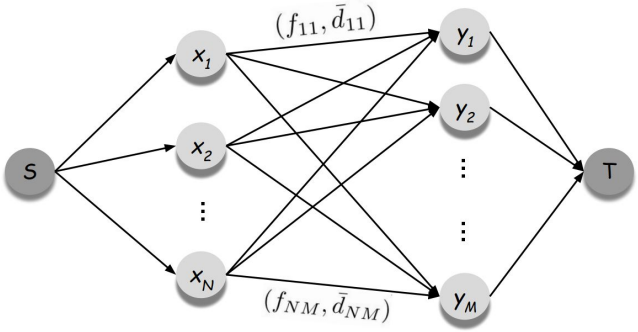
Being Linear B and Ancient Greek the same language with a different scripts, we can map their cognates.

- Jiaming Luo's Cognate Matching Framework was employed.
- Generative model: produces an Ancient Greek cognate for any Linear B word.
- Expectation-Maximization training
- Network-flow problem for cognate matching.

Cognates examples:

- ko-no-so → Κνωσός
- ko-wo → κόρος
- do-se → δώσει





Estimate Edit distance

Cognate Matching Model

Jiaming Luo outlined four language-agnostic cognate matching principles:

1. Distributional similarity of matching characters
2. Monotonic character mapping within cognates
3. Structural sparsity of cognate mapping
4. Significant cognate overlap within related languages

His model architecture deals with all these principles.
The machine learning model is responsible for the first two properties, the flow latent variable (F) accounts for the other two.

$$\mathbf{P}(\mathcal{X},\mathcal{Y}) = \sum_{\mathcal{F}\in\mathbb{F}} \mathbf{P}(\mathcal{F})\,\mathbf{P}(\mathcal{X}\mid\mathcal{F})\,\mathbf{P}(\mathcal{Y}\mid\mathcal{F},\mathcal{X}) \propto \sum_{\mathcal{F}\in\mathbb{F}} \mathbf{P}(\mathcal{Y}\mid\mathcal{X},\mathcal{F}) = \sum_{\mathcal{F}\in\mathbb{F}} \prod_{y_j\in\mathcal{Y}} \mathbf{P}(y_j\mid\mathcal{X},\mathcal{F})$$

$$\mathbf{P}(y_j\mid\mathcal{X},\mathcal{F}) = \sum_{x_i\in\mathcal{X}} f_{ij}\cdot\mathbf{P}_\theta(y_j\mid x_i)$$

Training Technique

Algorithm 1 Iterative training

Require:

$\mathcal{X}, \mathcal{Y}$: vocabularies,

T: number of iterations,

N: number of cognate pairs to identify.

$f_{i,j}^{(0)} \leftarrow \frac{N}{|\mathcal{X}| \cdot |\mathcal{Y}|}$ ▷ Initialize

for $\tau \leftarrow 1$ to T **do**

  $\theta^{(\tau)} \leftarrow \text{MLE-TRAIN}(f_{i,j}^{(\tau-1)})$

  $\bar{d}_{i,j}^{(\tau)} \leftarrow \text{EDIT-DIST}(x_i, y_j, \theta^{(\tau)})$

  $\tilde{f}_{i,j}^{(\tau)} \leftarrow \text{MIN-COST-FLOW}(\bar{d}_{i,j}^{(\tau)})$

  $f_{i,j}^{(\tau)} \leftarrow \gamma \cdot f_{i,j}^{(\tau-1)} + (1-\gamma) \cdot \tilde{f}_{i,j}^{(\tau)}$

  $\text{RESET}(\theta^{(\tau)})$

return $f_{i,j}^{(T)}$

function MLE-TRAIN($f_{i,j}^{(\tau)}$)

  $\theta^{(\tau)} \leftarrow \arg\max_\theta \prod_{y_j \in \mathcal{Y}} \Pr_\theta(y_j | \mathcal{X}, \mathcal{F})$

  return $\theta^{(\tau)}$

- Progressively updates latent flow values, used for loss computation
- Reinitialize the model at each E-step
- 3 output modes: MLE, Flow, and Flow with Expected Edits
- Distances gathered from multiple sampling + edit distance, or from character-wise probabilities product
- Flow Mode output: output of min-cost flow computed with the trained model

Dataset making

The Linear B cognate matching dataset employed was updated.
The dataset making comprised:

- Manual corrections to Luo's original dataset
- Brute-force greedy algorithm to associate Linear B words to possible Ancient Greek cognates from Homer (Iliad and Odyssey)
- Cognates refinement through structured prompt engineering
- Manual corrections
- Adding other matches from Tselentis lexicon (available only as a PDF, processed with OCR tools and manually revised).

Luo's original dataset: 919 Linear B words, 1418 Greek correspondences
Our dataset: 1911 Linear B words, 2383 Greek correspondences

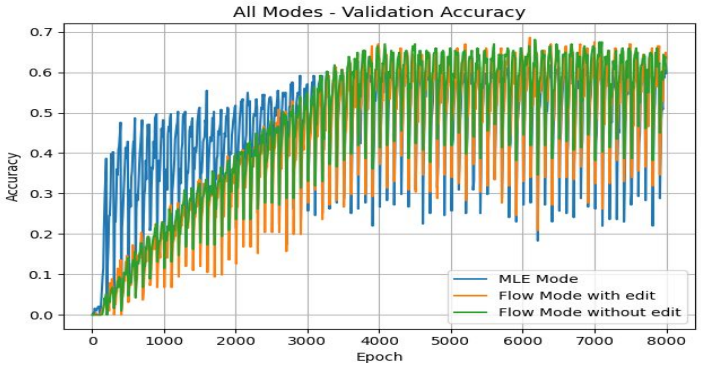
Cognate matching results

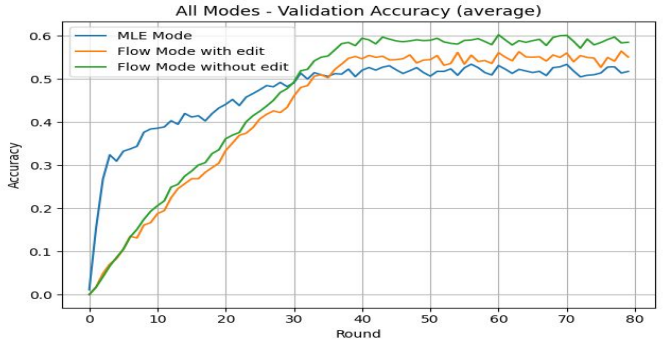
LSTM Init	Split	MLE	Flow w/o edit	Flow w/ edit
zero	Train	0.789	0.790	0.735
	Validation	0.663	0.674	0.696
	Test	0.761	0.761	0.728
custom	Train	0.411	0.423	0.435
	Validation	0.391	0.424	0.435
	Test	0.359	0.359	0.435

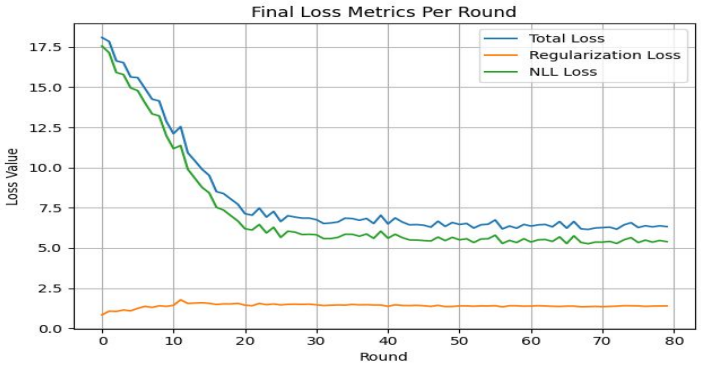
Table 3.4. Performance (Train/Validation/Test) on Luo's original dataset.

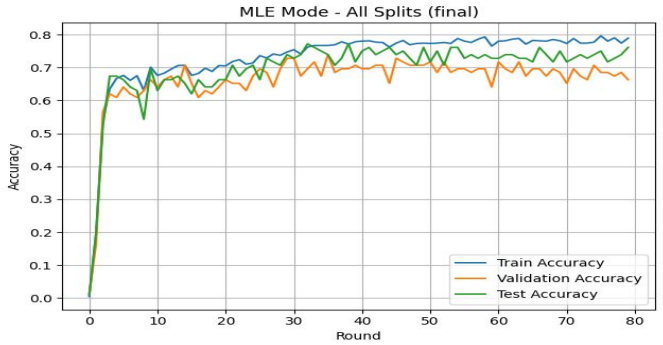
LSTM Init	Split	MLE	Flow w/o edit	Flow w/ edit
zero	Train	0.645	0.658	0.639
	Validation	0.597	0.602	0.634
	Test	0.625	0.646	0.646
custom	Train	0.238	0.275	0.367
	Validation	0.262	0.262	0.377
	Test	0.224	0.234	0.370

Table 3.6. Performance (Train/Validation/Test) on our dataset.









Auxiliary classifiers

Three additional tasks were identified to support translation:

1. Part of Speech Detection: noun, adjective, verb, adverb
2. Noun type: proper, toponym, ethnonym, common
3. Inflection Detection: thematic -o, thematic -a, athematic

Number	Case	Thematic *-o* (M.)	Thematic *-o* (N.)	Thematic *-a* (M.)	Thematic *-a* (F.)	Athematic (M./F.)	Athematic (N.)
Sing.	Nom.	*-o*	*-o*	*-a*	*-a*	variable	variable
Sing.	Gen.	*-ojo*	*-ojo*	*-ao*	*-a*	*-o*	*-o*
Sing.	Dat.	*-o*	*-o*	*-a*	*-a*	*-e/-i*	*-e/-i*
Sing.	Acc.	*-o*	*-o*	*-a*	*-a*	*-a*	variable (often = Nom.)
Plur.	Nom.	*-o/-oi*	*-a*	*-a*	*-a*	*-e*	*-a*
Plur.	Gen.	*-o*	*-o*	*-ao*	*-ao*	*-o*	*-o*
Plur.	Dat.	*-oi*	*-oi*	*-ai*	*-ai*	*-si/-ti*	*-si/-ti*
Plur.	Acc.	*-o*	*-a*	*-a*	*-a*	*-a/-e*	*-a*

POS Detection: clarifies words' role in the sentence.
NT: clarifies how nouns (and adjectives) should be interpreted
ID: explains how to interpret inflected terms thanks to declension table

Tested solutions

The first approach: training neural networks (BRNN, convolutional) on embedded tokenized Linear B words.
This approach demonstrated slow convergence and low performance.

The second approach: bag-of-words for the Linear B syllabograms and Tf-Idf features for n-grams.
This word preprocessing proved to be effective for all tasks, with a variety of models while also being much more efficient.
Due to the more satisfactory results, only this approach was tested using 5-fold Cross Validation with the following models:

1. Logistic Regression
2. Random Forest
3. Linear SVM
4. Multinomial Naive Bayes
5. Histogram-based Gradient Boosting Classification Tree
6. MLP classifier

Auxiliary tasks results: Linear SVM performs best in all tasks

Model	Accuracy	Macro-F1
Logistic Regression	**0.8848**	0.3826
Random Forest	0.8770	0.2745
Linear SVM	0.8845	**0.4129**
Multinomial Naive Bayes	0.8712	0.2439
HistGradientBoosting	0.8777	0.3390
Neural Network (MLP)	0.8790	0.3022

Part of Speech Detection

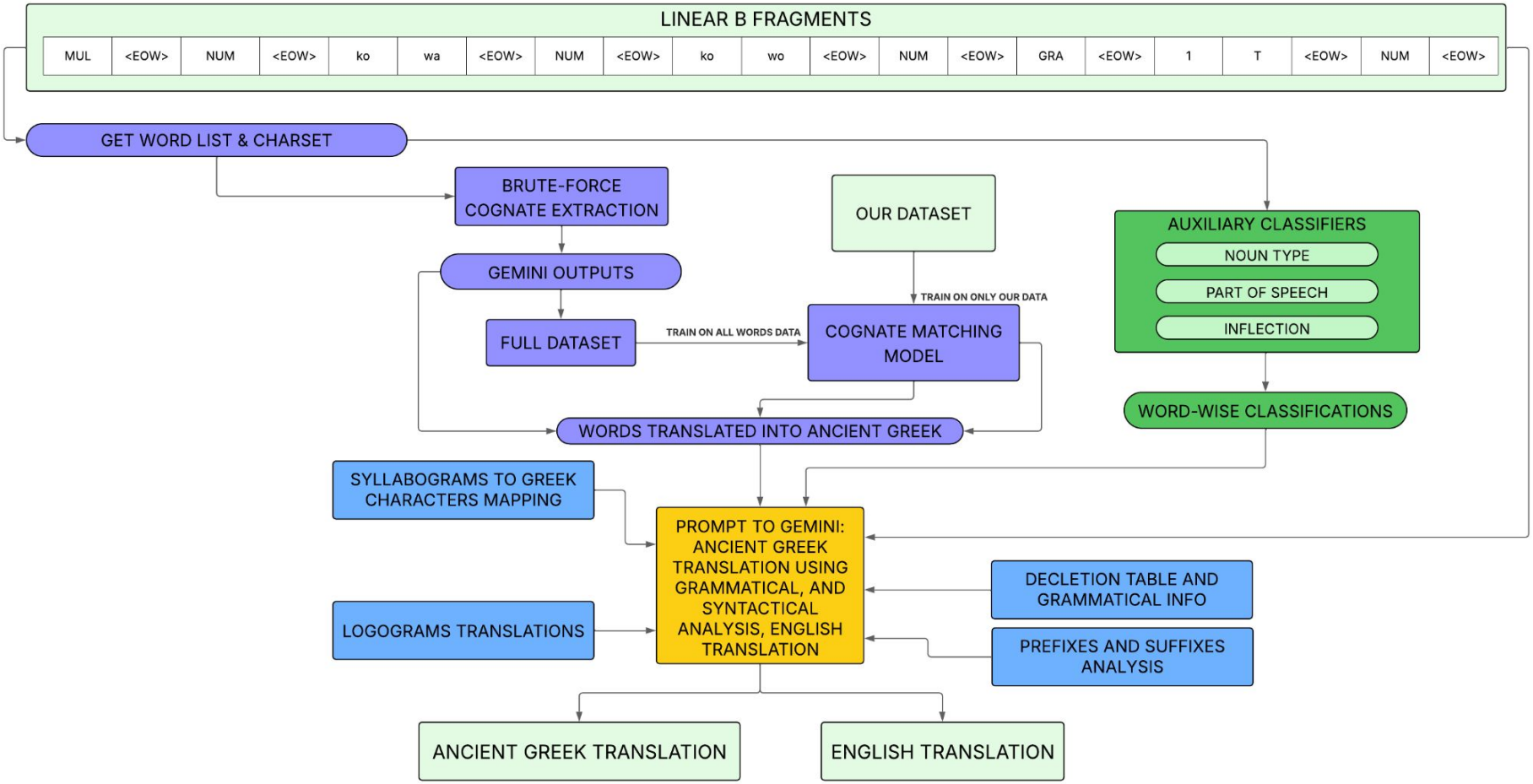
Model	Accuracy	Macro-F1
Logistic Regression	0.6341	0.4045
Random Forest	0.6015	0.2953
Linear SVM	**0.6401**	**0.4632**
Multinomial Naive Bayes	0.6044	0.2978
HistGradientBoosting	0.6012	0.3841
Neural Network (MLP)	0.6319	0.3653

Noun Type Classification

Model	Accuracy	Macro-F1
Logistic Regression	0.8097	0.7755
Random Forest	0.7405	0.6836
Linear SVM	**0.8372**	**0.8072**
Multinomial Naive Bayes	0.7298	0.6690
HistGradientBoosting	0.7855	0.7443
Neural Network (MLP)	0.7993	0.7611

Inflection Detection

Translation Pipeline



Results

12 documents, whose translation had been provided by Tselentis, have been manually evaluated.
Adjustments have been highlighted in red, mistakes in blue.

Document name: KN Ra 1540

Linear B text: *to-sa pa-ka-na PUG 50*
Greek translation: τόσα φάσγανα ΦΑΣΓΑΝΑ 50.
English translation: So many swords: 50 SWORDS.

Document name: KN So 4439

Linear B text: *a-mo-ta e-ri-ka te-mi-dwe-ta ROTA ZE 3 MO ROTA 1*
Greek translation: ἅρματα ἑλικά τερμιοέντα: τροχοί ζεύγη 3, μόνος τροχός 1.
English translation: Wheels made of willow tree, rimmed: 3 pairs of wheels, 1 single wheel.

Error Analysis & Evaluation

Overall, four main causes of error were identified in the analyzed documents:

1. insufficient/misleading cognate information;
2. misinterpretation of words' grammatical function (mostly due to wrong case selection);
3. auxiliary classifiers' mislabeling, often turning common names into proper names;
4. Logograms misinterpretation, especially for abbreviations.

Machine Translation metrics

1. BLEU
2. METEOR
3. ROUGE-(1, 2 and L)
4. TER
5. CHRF
6. WER

Metric	Greek	English
BLEU	0.641	0.768
METEOR	0.813	0.880
ROUGE-1 F1	0.818	0.903
ROUGE-2 F1	0.698	0.837
ROUGE-L F1	0.815	0.899
TER	0.187	0.158
CHRF	0.836	0.848
WER	0.189	0.161

Conclusions and future work

Overall, the LLM always manages to understand the context of the administrative Linear B records, and translates most of them appropriately.

Possible ideas for future work include:

- manual labeling of auxiliary classifiers' dataset;
- expanding auxiliary tasks (detecting tense/mood/voice for verbs, or case, gender and number for nouns/adjectives);
- fine-tuning an LLM using expert-validated Ancient Greek and English translations;
- forms normalization towards classical Ancient Greek;
- add possible interpretations for logographic abbreviations;
- evaluate the performance of this pipeline on other related languages.
- extending the systematic evaluation to more documents from the Linear B corpus