Written on Leaves or in Stones?: Computational Evidence for the Era of Authorship of Old Thai Prose

Attapol T. Rutherford [*]
Department of Linguistics
Faculty of Arts
Chulalongkorn University
attapol.t@chula.ac.th

Santhawat Thanyawong
Department of Linguistics
Faculty of Arts
Chulalongkorn University
santhawat.t@gmail.com

Abstract

We aim to provide computational evidence for the era of authorship of two important old Thai texts: *Traiphumikatha* and *Pumratchatham*. The era of authorship of these two books is still an ongoing debate among Thai literature scholars. Analysis of old Thai texts present a challenge for standard natural language processing techniques, due to the lack of corpora necessary for building old Thai word and syllable segmentation. We propose an accurate and interpretable model to classify each segment as one of the three eras of authorship (Sukhothai, Ayuddhya, or Rattanakosin) without sophisticated linguistic preprocessing. Contrary to previous hypotheses, our model suggests that both books were written during the Sukhothai era. Moreover, the second half of the *Pumratchtham* is uncharacteristic of the Sukhothai era, which may have confounded literary scholars in the past. Further, our model reveals that the most indicative linguistic changes stem from unidirectional grammaticalized words and polyfunctional words, which show up as most dominant features in the model.

1 Introduction

The time periods of authorship for many of the old Thai texts are still being disputed and debated, as the identities of the authors are not always well established. Previous approaches often require diachronic close reading of the text to identify the key elements of style or specific linguistic changes that characterize the writing of the era. Such analysis is limited to qualitative accounts drawn from hand-selected textual evidence. In this work, we build a model that infers the time period of authorship for old Thai prose and reveals diachronic linguistic changes while tolerating the natural language processing (NLP) resources and corpora.

Computational approaches analyzing semantic change in old Thai text face many critical challenges due to poverty of NLP resources. The field lacks texts that could serve as representative examples from each era because solid historical evidence can identify the time of writing for only a few texts. Old Thai prose is especially rare. Consequently, we do not have enough texts to re-train syllable and word segmenters or fit classification models. The currently available Thai syllable and word segmentation algorithms do not perform well on old Thai text, owing to dramatically different orthography and vocabulary. Worse still, some representative Thai texts are significantly damaged inscriptions on stones, which impede sentence-level or even word-level analysis. Thus, to analyze old Thai prose, we cannot rely on automatic syllable and word segmentation, nor on models that require large amounts of data from the same era.

In this work, we propose an accurate and interpretable classification model for analyzing the time period of authorship from textual segments of old Thai prose from *Traiphumikatha* (ไตรภูมิกถา) and *Pumratchatham* (ปูมราชธรรม), whose time of authorship is still debated. Unlike most author attribution models, our model scans through and operates at the text segment level; hence the name Maximum Entropy Searchlight model. The model uses varying-length character n-grams as features to classify textual segments into one of the eras. We shrink the model coefficients to reveal the character n-grams that are distinguishing linguistic features of each era. The model spotlights specific text segments that are characteristic of the era where the book was written and provides computational evidence of the era of authorship for the books in question.

The main contribution of this work can be summarized as follows:

[*] corresponding author

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 81–85
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

- We propose an accurate and interpretable model for identifying the era of authorship of old Thai prose. The model classifies text segments with high accuracy, reveals some of the linguistic changes from the Sukhothai to the Ayuddhya era, and serves as a visualization tool for further linguistic analysis.
- We are the first to provide statistical evidence that *Traiphumikatha* and *Pumratchatham* might be both written in the Sukhothai era, contrary to previous hypotheses.
- As a more general principle, we conclude that grammaticalized words and polyfunctionalized words are the strongest distinguishing indicators of prose from the Sukhothai era.

2 Background and Related Work

In diachronic studies, Thai language eras are roughly divided by historical timeline of state establishment: Sukhothai (1249-1438), Ayuddhya (1350-1767), Thonburi (1767-1782), and Rattanakosin (1767-present). Ayuddhya and Rattanakosin eras are sometimes further divided into 'early,' 'mid,' and ‘late,' depending on the individual research purposes. Due to the gradually changing nature of languages, a language change can be observed only when the language samples in comparison are taken from quite distant eras. It is widely believed that *Traiphumikatha* was written in Sukhothai era although the oldest copy was found in Thonburi era and the proof of era of authorship was never rigorously established (Eawsriwong, 1982). *Pumratchatham* is believed to be written during late Ayuddhya (1688-1767) as the orthography and letter types appear on the first page were usually found in late Ayuddhya books.

Our task can be seen as an author attribution problem or style-change detection problem. These models have utilized all levels of features: lexical, character, syntactic, discourse, and structural (Stamatatos, 2009; Ferracane et al., 2017). Various neural network architectures have been explored in the context of this task (Shrestha et al., 2017). Yet, our task differs in that each class has a mixture of authors. We want to use feature-based models for their interpretability, plus want the model to be accurate at the level of small text segments.

3 Data and Model Descriptions

The reference ground truth texts for each era are: stone inscriptions (Sukhothai era), *Historical Archive on Kosapan's trip to France* (จดหมายเหตุโกษาปานไปฝรั่งเศส) (Ayuddhya era), and *Historical Archive on Luang Udomsombat* (จดหมายเหตุหลวงอุดมสมบัติ) (Rattakosin era). The stone inscriptions vary in their quality, as some are broken stone fragments and not full texts. The identity of the authors of these inscriptions is either unknown or disputed. *Traiphumikatha* and *Pumratchatham* are the two texts whose time of authorship we want to investigate. We use the manually cleaned version of the texts used by literary scholars because different orthography could bias the models. The data sizes and the class distribution are shown in Table 2.

Text collection	Character count	Segment count
Ground truth		
Sukhothai era	39,700	873
Ayuddhya era	39,872	984
Rattanakosin era	411,134	10,182
Text in question		
Pumratchatham	110,118	2,741
Traiphumikatha	349,162	8,484

Table 1: Data statistics of the five text collections

Our goal is to create a three-way (Sukhothai vs Ayuddhya vs Rattanakosin) classification model that is accurate enough to give us statistical evidence for time of authorship, and interpretable enough to reveal linguistic changes that might require further analysis at the small segment level. We propose Maximum Entropy Searchlight model, which is a multi-class logistic regression model (or Maximum Entropy model) with bag of varying-length character n-gram features and an L1 penalty (Tibshirani, 1996). We formulate the task as text segment classification, with each text divided into non-overlapping contiguous character segments. Numerals, indentation, and punctuations serve as segment dividers, but we cap the segment length to be at most 40 characters, which is right around the median segment lengths.

The model scans through each substring of each segment like a searchlight sweeping across the text, hence the name of the model. The L1 penalty acts as a feature selection mechanism to restrain the model to keep only a handful of interpretable features, while shrinking the rest to zero. Since our model is saturated with both redundant and unhelpful features, this penalty is suitable.

Should we use fixed-length n-grams or varying-length n-grams? We run 10-fold cross-validation to compute the accuracy rates of fixed-

Crossvalidated accuracy		n-gram min	n-gram max	Params	Non-zero params	
0.99	±0.004	2	6	529k	1190	±16
0.98	±0.005	3	6	524k	1278	±24
0.98	±0.008	4	6	487k	2027	±24
0.96	±0.008	5	6	387k	2956	±49
0.98	±0.005	2	2	4k	1079	±14
0.98	±0.006	3	3	37k	1109	±13
0.97	±0.007	4	4	99k	1727	±17
0.96	±0.008	5	5	166k	2477	±33
0.94	±0.012	6	6	221k	3229	±24

Table 2: Varying-length n-gram features perform the best while keeping the number of non-zero parameters relatively low.

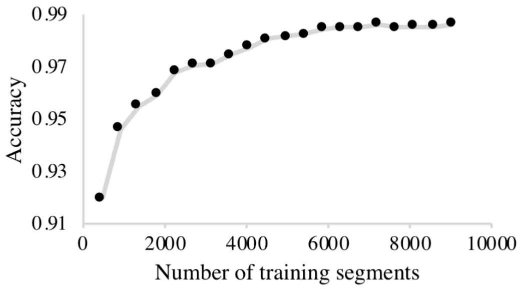

Figure 1: The classifier requires only a small portion of the books to be able to classify the rest at high accuracy.

length n-gram models and varying-length n-gram models ($n \in [2, 6]$). The varying-length n-gram models outperform the best fixed-length models although the L1 penalty shrinks the number of parameters of both types of models to be quite similar (Table 2). The best model only requires (non-zero) 1190 parameters. Our results suggest that varying-length n-gram features are more effective than fixed-length n-gram features even when the number of the parameters are comparable.

Is the model accurate enough to use for unknown texts? We vary the amount of training data from around 4% (454 segments) to 75% (9029 segments) and test the model on the test set, which constitutes the remaining 25% of each book. The final model uses varying-length character n-grams with $n \in [2, 6]$, without fitting the intercepts. The accuracy of the model grows logarithmically with the amount of training data, like a typical learning curve of a classifier (Figure 1). Strikingly, the model requires only 40% (4815 segments) of the text from each era to achieve 98% accuracy (Table 3). This low training fraction suggests that the style of writing varies substantially across eras, because the model can capture most of the variation with substantially fewer samples than available. This result also suggests that we can readily apply this model on texts whose era of authorship is unknown.

Era	Precision	Recall	F1
Sukhothai	0.96	0.85	0.90
Ayuddhya	0.98	0.95	0.97
Rattanakosin	0.99	0.99	0.99
Macro average	0.98	0.93	0.95
Micro average	0.98	0.98	0.98

Table 3: Classification results based on the best cross-validated model

Does the model present interpretable results? We examine the 30 most salient model coefficients (weights) for linguistic changes. For the Sukhothai era, 15 of those features correspond to known changes studied in Thai historical linguistics. Examples include /lɛ́:w/ and /jù:/ (Sriprasit, 2003), /pen/ (Jaratjarungkiat, 2012), /thɯ̌ŋ/ and /thɤ̌ŋ/ (Rodphan, 2012), /mí/ and /bɔ̀mí/ (Jampathip, 2014), and /ʔân/ /nân/ and /nán/ (Suwangphanich, 2017). This correspondence demonstrates how our model can pinpoint specific words for further linguistic analyses.

4 When were *Traiphumikatha* and *Pumratchatham* written?

We classify each 40-character segment of the text and gather the computational evidence for the era of authorship. For each of the two books, we compute the distribution of eras as classified by the model, along with the total log-likelihood of each era given the model. We also compute the distribution of high-confidence classifications for each era, where the score exceeds 0.9. 46% and 41% of the segments from Traiphumikatha and Pumratchatham, respectively, pass this 0.9 threshold (Table 4). The model excludes the intercept terms, to avoid biasing the classification.

Our model supports the hypothesis that *Pumratchatham* was written in the Sukhothai era, contrary to what is popularly believed. 66.8% and 57% of the segments from *Traiphumikatha* and *Pumratchatham* respectively are classified as more similar to the stone inscriptions from the Sukhothai era. Many scholars have hypothesized that *Traiphumikatha* might be written in the Ayuddhya era. Surprisingly, our model gives very little evidence to support this hypothesis, as less than 5% of the segments are classified as Ayuddhya.

The Maximum Entropy Searchlight model vi-

Era	Traiphumikatha					Pumratchatham				
	Classification distribution		>0.9 only distribution		Total likelihood	Classification distribution		>0.9 only distribution		Total likelihood
Sukhothai	5664	67%	2947	75%	-7984	1566	57%	690	58%	-3554
Ayuddhya	286	3%	19	0%	-43511	60	2%	3	0%	-14982
Rattanakosin	2533	30%	956	24%	-24528	1115	41%	498	42%	-5640

Table 4: The distribution of classified segments and the total likelihood suggest that Traiphumikatha and Pumratchatham were likely written in the Sukhothai era, contrary to previous hypotheses.

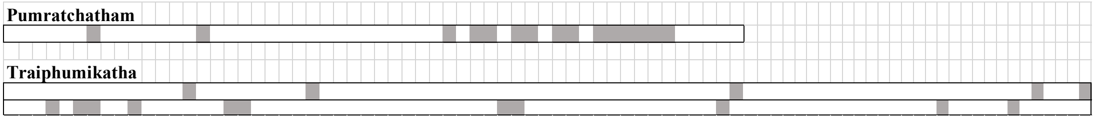

Figure 2: The language of the second half of *Pumratchatham* does not resemble the language from the Sukhothai era. 30-segment blocks are shaded if the majority of its 40-character segments are classified as Rattanakosin era, while the unshaded blocks are Sukhothai.

sualizes potential style changes within the book and spotlights the regions that deserve further investigation. We group 40-character text segments into a blocks and visualize the majority class for each block (Figure 2). It turns out that the non-Sukhothai parts of *Pumratchatham* are clustered towards the end of the book, while non-Sukhothai parts are distributed more uniformly in Traiphumikatha. The era of authorship of this book may be more contested for this reason.

5 Grammaticalization and Polyfunctionalization across Eras

Some of the most common features are words that undergo the process of grammaticalization over time such as /lɛ́:w/, /jù:/, and /pen/. Grammaticalization refers to the phenomenon where a lexical item becomes a grammatical marker and develops new grammatical functions (Hopper and Traugott, 2003). Grammaticalization is unidirectional in the sense that grammatical forms and markers cannot become lexical again. This implies that the linguistic characteristics of a grammaticalized word are different in each stage of changes. Thus, grammaticalized words can strongly characterize eras.

Polyfunctional words (words that can take multiple part of speech tags) form another group of linguistic changes indicative of eras of authorship. We found 6 words to be polyfunctional observable in synchronic Thai grammar. These are /sǐ:a/, as verb and completive aspect marker (Iwasaki et al., 2005), /hɛ̀ŋ/ and /khâ:ŋ/, as noun and preposition, /thâw/, as noun and adverb, /bâ:ŋ/, as pronoun and adverb (Royal Institute dictionary B.E. 2554) and /ʔɔ̀:k/, as verb and adverb (Wongsri, 2004). Polyfunctionality of a word can be seen as a synchronic product of the unidirectional grammaticalization process called `layering', which is the persistence of older forms and meanings alongside newer ones (Hopper and Traugott, 2003) . Our model reveals this synchronic state of grammaticalization and unidirectional linguistic changes that characterize the differences across the eras.

In sum, 15 of 30 extracted words given by the model can be best explained in a single theme of unidirectionality of change, a tendency that forms the backbone of grammaticalization (diachronic change) and layering (synchronic resultant state of the change). Thus, these words, along with grammaticalization perspective, can best validate the Maximum Entropy Searchlight Model as a tool to provide the statistical evidence for the era of authorship.

6 Conclusion

We present the Maximum Entropy Searchlight model, an accurate and interpretable model for identifying the era of authorship of old Thai prose. The model lends reliable computational evidence for the era of authorship because it can classify the era of the ground truth text collections at almost perfect accuracy. In addition, the model can shed light on each individual segment to discover specific linguistic changes that are important indicators for each era. These attributes not only speed up the process of qualitative linguistic analysis, but also reveal an overarching theme of unidirectional grammaticalization, characterizing the differences across the eras.

Acknowledgements

This research is funded by Grants for Development of New Faculty Staff at Chulalongkorn University.

References

Nidhi Eawsriwong. 1982. หลักฐานทางประวัติศาสตร์ในประเทศไทย [Historical Evidence in Thailand]. Bhannakij Trading.

Elisa Ferracane, Su Wang, and Raymond Mooney. 2017. Leveraging discourse information effectively for authorship attribution. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 584--593.

Paul J Hopper and Elizabeth Closs Traugott. 2003. *Grammaticalization*. Cambridge University Press.

Shoichi Iwasaki, Preeya Ingkaphirom, and Inkapiromu Puriyā Horie. 2005. *A reference grammar of Thai*. Cambridge University Press.

Nida Jampathip. 2014. *The development of the negators "bo" "mi" "pai" "mai" in Thai*. Ph.D. thesis, Chulalongkorn University.

Sureenate Jaratjarungkiat. 2012. *The development of the word /pen/ in Thai*. Ph.D. thesis, Chulalongkorn University.

Krongkan Rodphan. 2012. /thɯ̌ŋ/: a historical study. Master's thesis, Chulalongkorn University.

Prasha Shrestha, Sebastian Sierra, Fabio Gonzalez, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 669--674.

Mingmit Sriprasit. 2003. A diachronic study of / læǽw/, /yuù/ and /yuùlæǽw/. Master's thesis, Chulalongkorn University.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538--556.

Wasitthi Suwangphanich. 2017. Development of demonstratives in thai. Master's thesis, Chulalongkorn University.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267--288.

Katbandit Wongsri. 2004. A semantic network of / ʔɔɔk/ in thai: a cognitive semantic study. Master's thesis, Chulalongkorn University.