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

Attapol T. Rutherford attapol.t@chula.ac.th

Santhawat Thanyawong santhawat.t@gmail.com

Department of Linguistics, Faculty of Arts, Chulalongkorn University

What we learned

- ***Grammaticalized words and polyfunctional words are key indicators of era of authorship in old Thai***
- ***Word segmentation of old Thai Text is not necessary for the analysis of era of authorship.***

Problems

- The identification of era of authorship in old Thai prose lacks computational/quantitative evidence.
- Old Thai text breaks standard NLP tools due to linguistic differences.

Our approach

- Character-ngram Maximum Entropy model

Data and model description

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

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.

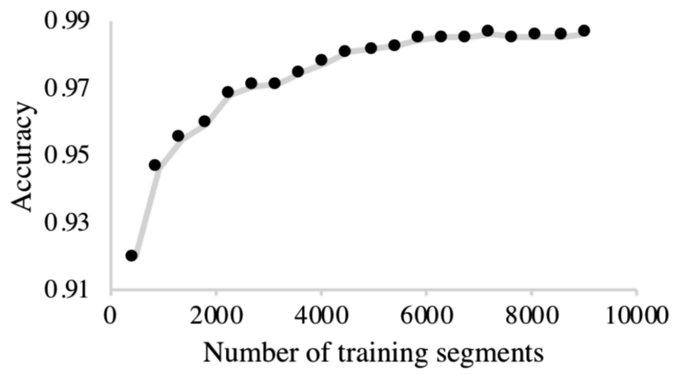

Figure 1: The classifier requires only a small portion of the books to be able to classify the rest at high accuracy.

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.

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

This result suggests that we can readily apply this model on texts whose era of authorship is unknown.

When were Traiphumikatha and Pumratchatham written?

Era	Traiphumikatha Classification distribution		Traiphumikatha >0.9 only distribution		Traiphumikatha Total likelihood	Pumratchatham Classification distribution		Pumratchatham >0.9 only distribution		Pumratchatham 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.

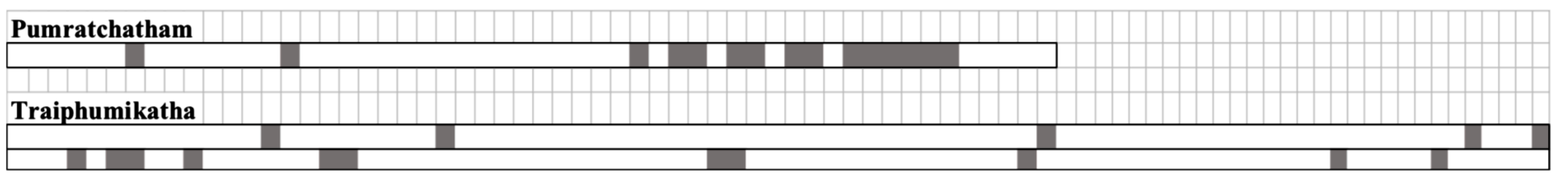

Figure 2: The language of the second half of *Pumratchatham* does not resemble the language from the Sukhothai era. 30-segment blocks are shaded if the majority of its 40-character segments are classified as Rattanakosin era, while the unshaded blocks are Sukhothai.

- *Our model supports the hypothesis that Pumratchatham was written in the Sukhothai era, contrary to what is popularly believed.*
- *Many scholars have hypothesized that Traiphumikatha might be written in the Ayuddhya era. Surprisingly, our model gives very little evidence to support this hypothesis*