



A Linguistic Study of Thai International News Headlines Based on Quantitative Linguistics

Shi Chen*, Mingyao Huang

Guangdong University of Foreign Studies, Guangzhou, Guangdong, China.

How to cite this paper: Shi Chen, Mingyao Huang. (2024) A Linguistic Study of Thai International News Headlines Based on Quantitative Linguistics. *Journal of Humanities, Arts and Social Science*, 8(4), 953-958. DOI: 10.26855/jhass.2024.04.024

Received: March 31, 2024

Accepted: April 30, 2024

Published: May 29, 2024

***Corresponding author:** Shi Chen, Guangdong University of Foreign Studies, Guangzhou, Guangdong, China.

Abstract

This article compiles news materials from Thailand's mainstream media from August 2015 to August 2021, including Public Opinion, Frontline, Daily News, Thai Lat News, New News, Siam Politics, BBC (Thai), National Newspaper, and the Thai Post. A total of 650,039 news titles were collected. Through classification tag screening, 8620 international news articles were obtained, accounting for 1.33%. After completing the collection of the news corpus, the author performed word segmentation by using the Python library of PyThaiNLP. Through automatic word segmentation and manual review, a Thai news headline corpus was finally obtained with a sample size of 8620 headlines, a total of 6076 types, and a total of 176,512 tokens. From the analysis of the subject words, it can be inferred that the COVID-19 epidemic has been the primary focus of Thailand's international news in recent years. The word COVID-19 ranks first with a frequency of 1,429. In addition, there are other manifestations of COVID-19, with a total frequency of 2,292. The popularity of this event surpasses that of all others. In addition, Myanmar's military issues are also receiving significant attention in international news coverage in Thailand.

Keywords

Quantitative linguistics, Thai language, news headlines

1. Introduction

The development of the Internet has brought audiences into the era of topic reading. After traditional news media entered the Internet, they faced various communication platforms and made changes in the expression of news content. Thai news media have also entered a new stage of headline reading. Thai news headlines are unique in that they often use abbreviations, colloquialisms, slang, metaphors, references, incomplete grammatical structures, and many other characteristics. Therefore, non-native Thai readers have difficulty understanding Thai titles. The study of international news headlines in Thai aims to sort out Thailand's reports on news from various countries and analyze the number of news reports, hot events, high-frequency vocabulary, etc. in each country, which will help to examine the spread of each country's international image in Thailand from the perspective of others. Among them also includes China.

At present, interdisciplinary research has emerged in title language research based on traditional humanities and social science research. Communication, journalism, and linguistics have begun to try to integrate with computer science. By using corpus, LIWC, and other statistical software for data mining and analysis, title language research is no longer mainly based on static description, and the research process has become more intuitive. With the support of data, the research results are more scientific; computer software and computer applications are also combined with actual needs to continuously optimize algorithms and other computer application technologies to assist title language research (Teewara Saengin, 2020). Therefore, this article aims to build a Thai news corpus, collect Thai international

news from August 2015 to August 2021, and use natural language processing technology to preprocess the corpus and process it according to the relevant measurement indicators of quantitative linguistics. The collected corpus is analyzed and finally linguistically interpreted in order to outline the hot spots and trends in Thai international news, thereby analyzing China's international image and discourse spread in Thailand.

2. Literature Review

Domestic-related research on Thai news has a long history. In 1985, Wang Rongjiu published the study "Overview of Thai Journalism" in the international press, and research on Thai news media began to enter the domestic academic community. Among the many media studies in Thailand, most studies are on Thai Chinese media. In recent years, related research on the analysis of China's image in Thailand's mainstream media and new media has gradually become a research hotspot. Including focusing on the reporting of individual major news events in Thailand, such as "Critical Discourse Analysis of Thai Media Reports on Chinese Tourists" (Huang Tingting, 2017) and "Analysis of News Coverage of the "Belt and Road Initiative" in Thai English Media from the Perspective of Evaluation Theory" (Gu Dongxue, 2019), etc. In addition, the title characteristics of Thai mainstream media are studied from a linguistic perspective. For example: Le Yufei's "Research on the Characteristics of Thai Rath Newspaper News Titles" (Le Yufe, 2018) and Zhang Junjie's "Research on the Linguistic Characteristics of Thai Media News Titles in Multiple Contexts" (Zhang Junjie, 2021), etc.

At the same time, Thai scholars have gradually carried out research on the language of Chinese news headlines, keeping up with international trends and keeping close to the dynamics of Chinese society. Chuanpit Thiamtanom published "Analysis of News Titles of "People's Daily" and "Xinhua Daily" (Chuanpit Thiamtanom, 2014) in the *Journal of Chinese Studies*, analyzing the linguistic characteristics of the two newspapers from the meaning of words and the grammatical structure of phrases and sentences. Nednamthip Buddawong published "Political Discourse on COVID-19 Vaccine in Chinese News Headlines" (Nednamthip Buddawong, 2022) published in the *Ubon Ratchathani University Arts Journal*. He follows current events and uses critical analysis discourse theory to explore the political discourse in news slogans related to the new coronavirus epidemic. Kulnapa Siridissaku's study "China's National Image in Thai Mainstream Media (2013-2019)" (Kulnapa Siridissaku, 2021) examines the portrayal of China in three Thai mainstream media outlets: "Thai Daily", "Daily News" and "Khaosod." The study utilizes content analysis to investigate how Thailand's mainstream media influences China's national image through news selection and construction. Through a comparative analysis, it elucidates the factors influencing Thailand's mainstream media in shaping China's national image. Additionally, the study includes a questionnaire survey to gauge the extent to which local Thai audiences relate to the Chinese national image as portrayed by the mainstream media.

In summary, it can be seen that the current research on Thai news language headlines at home and abroad mainly adopts the traditional linguistic paradigm, and does not take large-scale Thai news headline corpus as the research object, and conduct research and analysis through relevant indicators of quantitative linguistics.

3. Corpus sources and corpus processing

The corpus of this article comes from the Thai news corpus constructed by the Guangzhou Non-General Language Intelligent Processing Laboratory. The corpus is managed by the School of Information Science and Technology/Cyberspace Security of Guangdong University of Foreign Studies. It has collected more than 800,000 Thai mainstream media news. This article selects news materials from Thailand's mainstream media from August 2015 to August 2021, including Public Opinion, Frontline, Daily News, Thai Lat News, New News, Siam Politics, BBC (Thai), National newspaper and the Thailand Post, a total of 650,039 news titles were obtained. Through classification and label screening, 8620 international news articles were obtained, accounting for 1.33%.

After completing the collection of the news corpus, the author performed word segmentation by importing the Python library of PyThaiNLP. The Python library has multiple Thai word segmentation algorithms. This article chose the whitespace+newline algorithm. Through automatic word segmentation and manual review, a Thai news headline corpus was finally obtained with a sample size of 8620 headlines, a total of 6076 Types, and a total of 176,512 Tokens.

4. Quantitative analysis

For quantitative analysis, this article uses QUITA software, whose full name is Quantitative Index Text Analyzer, developed by Palacki University in the Czech Republic. Through QUITA software, this article conducts quantitative

analysis on h-point, type ratio (TTR), vocabulary richness (R1), repetition rate (RR) and entropy (Entropy).

4.1 High-frequency word statistics: h-point measurement

Point h is a critical point in the rank frequency distribution of words in the text. The rank-frequency distribution of words is to arrange each word (type symbol) in the text in descending order of its frequency and number it in sequence from 1 to v. Each order r corresponds to a frequency value f(r). The so-called h point is the point where $r=f(r)$ on the rank frequency distribution of words. The words before point h all satisfy $r<f(r)$, and the words after point h all satisfy $r>f(r)$. In many rank-frequency distributions, it may not be possible to find a word that exactly satisfies $r=f(r)$. In this case, point h is actually between two adjacent words in frequency order. Assume that their frequency sequences are r_1 and r_2 ($r_2>r_1$), then there must be $r_1<f(r_1)$ and $r_2>f(r_2)$. The intersection of the straight line passing through the two points $(r_1, f(r_1))$ and $(r_2, f(r_2))$ and the straight line $y=x$ is point h. Point H divides the word frequency curve into two parts: particles - syntactic meaning words and substantive meaning words. Parts (can be roughly understood as two parts, function words and content words), in the range $[1, h]$ are the function word parts. Compared with the number of word types (V) in the text, although they are not large in number, they are in the text. The frequency of use has an absolute advantage. There are a large number of content words in the range of $[h, V]$. Although their frequency of use is not high in the text, they have made a great contribution to the richness of vocabulary. According to calculations, this time, the h point in the corpus text studied is 158.5.

4.2 Vocabulary diversity: type ratio (TTR), vocabulary richness (R1), repetition rate (RR), relative repetition rate (RRmc), and entropy (Entropy) calculations

According to statistics, the total number of Types in the target corpus is 6076, the total number of Tokens is 176512, the Type-to-Token Ratio (TTR) is Types/Tokens, and the calculated TTR= 0.034423.

R1 is the vocabulary richness index. It is an estimate of the proportion of content words (examples) in the text based on the h point. To calculate vocabulary richness, we first need to obtain the corresponding cumulative frequency distribution $F(r)$ based on the rank frequency distribution $f(r)$ of words in the text. $F(r)$ is the cumulative probability of words with frequency order from 1 to r, which is the proportion of the total frequency of these words to the total number of text examples. According to calculations, the R1 value in the corpus text of this study is 0.350855.

The indicator RR refers to the repetition rate. Like entropy, the metric RR is calculated based on the occurrence probability of words in the text. Suppose the occurrence probability of any word in the text is pr (that is, the ratio of its frequency to N), then $RR = \sum_{r=1}^v (r-1)^{-v} \left[\sum_{p=1}^r pr^2 \right]$. The repetition rate also reflects the richness of words in the text, but its value is inversely proportional to the richness of words. According to calculations, the repetition rate (RR) value in the corpus text of this study is 0.011633.

Entropy is a statistic used to describe diversity, inconsistency, and uncertainty. The entropy of word frequency: $H = -\sum_{i=1}^v (f_i/N) \log_2(f_i/N)$. In word frequency, the smaller the entropy, the more words are concentrated in a certain part, and the greater the entropy. The larger it is, the more evenly the words are distributed. According to calculations, the entropy value in the corpus text of this study is 8.470767.

After completing the calculation of the above measurement indicators, the author performed distribution fitting through the Altmann-Fitter software and found that its fitting effect with Zipf-Alekseev and Zipf-Mandelbrot law is good.

In the Parameters box of Altmann-Fitter software shows the values of several parameters when the selected distribution best fits the data set. X^2 is the chi-square value, $P(X^2)$ is its probability, and DF is the degrees of freedom of the chi-square test. $P(X^2)$ values much greater than 0.05 indicate excellent fitting results. C is the difference coefficient, which is a function of the chi-square value ($C=X^2/N$), where N is the total number of observed objects in the data set). When the size of the data set is too large and the chi-square test fails, C can be used to judge the quality of the fitting effect (generally speaking, when $C<0.02$, it means that the fitting effect is good, and when $C<0.01$, it means that the fitting effect is very good). R2 is the coefficient of determination of fitting. Judging from the fitting effect with the Zipf-Alekseev distribution, the $P(X^2)$ value is 0.4269, which is much greater than 0.05, and the C value is 0.004, which is far less than 0.01. The fitting effect is good.

Judging from the fitting effect with the Zipf-Alekseev distribution, the $P(X^2)$ value is 0.1489, which is much greater than 0.05, and the C value is 0.0044, which is far less than 0.01. The fitting effect is also good.

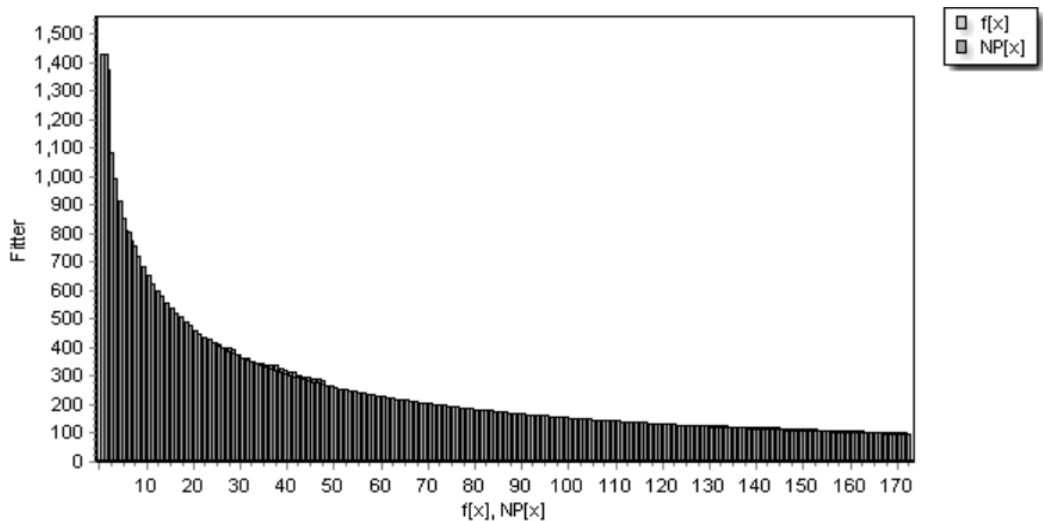


Figure 1. Zipf-Alekseev histogram distribution.

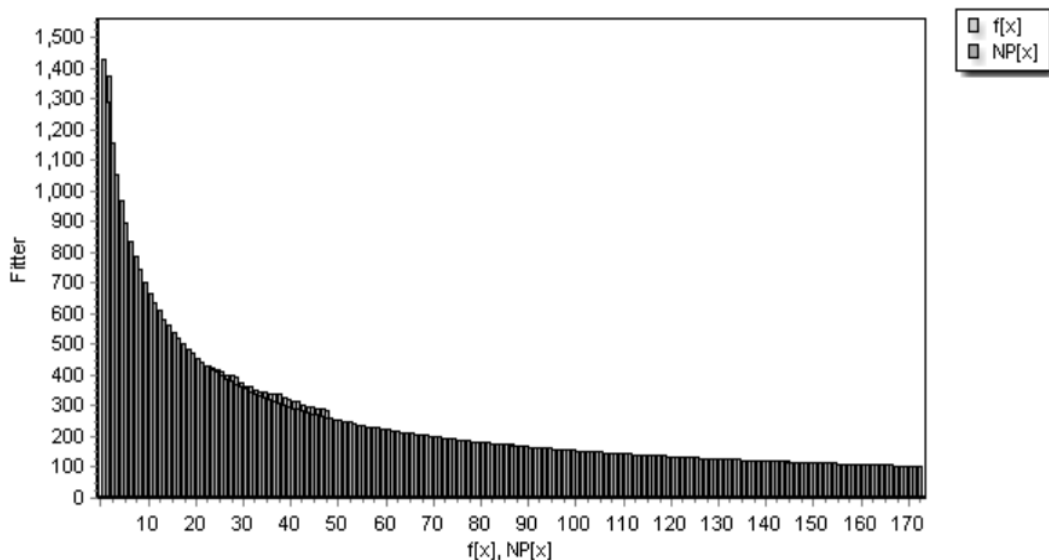


Figure 2. Zipf-Mandelbrot histogram distribution.

5. Linguistic explanation

According to the h-point value calculated previously, which is 158.5, the author organized the top 99 high-frequency words as follows and translated the content words. According to statistics, there are a total of 97 words before point h, of which 30 are notional words, accounting for 30.9%.

By analyzing the frequency of occurrences of country names in the above table, we can see that the word "China" appears 628 times, ranking 10th, and the word "United States" appears 453 times, ranking 19th. However, in reality, "United States" has many variations in Thai newspapers. In addition to the common สหรัฐ, there is also เม็กซิโก, which also refers to the United States. The frequency of the word is 326, the order is 39, and the combined frequency of the two words is 779, exceeding "The frequency of occurrence of the word "China". Generally speaking, the expression of สหรัฐ is more standardized and mostly appears in serious news. The expression of เม็กซิโก is more colloquial and often appears in non-serious news or popular entertainment newspapers. Although the combined frequency of the two expressions in the United States exceeds that of China, "Hong Kong, China", as a region that attracts much attention from Thailand, has a frequency of 180 and an order of 80. The combined frequency of the two expressions in China

is 808, slightly more than that of the United States. frequency. In addition, keywords related to the United States also include the words "Trump" and "Biden". Among them, the frequency of the word "Trump" is 291, the order is 46, and the frequency of "Biden" is 195, the order is 67. In addition to China and the United States, the frequency of the word "Myanmar" is 515, and the order is 14. It can be seen that the attention to the Myanmar issue in Thailand's international news cannot be ignored. In addition, the frequency of "India" is 304 and the order is 43, the frequency of "UK" is 283 and the order is 48, the frequency of "Japan" is 235 and the order is 54, and the frequency of "Russia" is 175. The order is 86. As a territorial country of Thailand, Myanmar has had constant disputes with Thailand in history. In recent years, Myanmar has been experiencing constant political problems, which have brought about problems such as drug trade, smuggling into the country, human trafficking, and epidemic prevention and control. Therefore, it has attracted great attention from Thailand, and it also makes sense.

Table 1. High-Frequency Notional Words

Rank	High-frequency words	Notional words translation	Frequency	Rank	high frequency words	Notional words translation	Frequency
1	โควิด	covid	1429	54	ญี่ปุ่น	Japan	235
4	โค	covid	863	58	ทหาร	army	235
6	วิด	pneumonia	814	67	ไบเดน	Biden	195
10	จีน	China	628	72	ป่วย	Sick	187
12	โลก	Word	576	74	ถล่ม	Attack	186
13	ไทย	Thailand	551	75	ประเทศ	country	184
14	เมียนมา	Myanmar	515	76	ปธน.	President	184
16	วัคซีน	vaccine	506	77	ล็อก	blockade	182
19	สหรัฐ	United States	453	80	ฮ่องกง	Hongkong	180
26	ตาย	death	409	83	คุม	control	177
32	ติดเชื้อ	Infect	365	86	รัสเซีย	Russia	175
39	มะกัน	America	326	88	คาวิน	curfew	168
43	อินเดีย	India	304	89	ชาติ	country	168
46	ทรัมป์	Trump	291	90	แกะรอย	track	167
48	อังกฤษ	U.K.	283	91	ช่วย	help	166
49	ฉีดวัคซีน	Vaccination	254	96	ระบาด	Spread	160
50	ทั่วโลก	worldwide	251	99	ผู้ป่วย	patient	155
51	ด้าน	Combat	248	99	รัฐ	Nation	153
52	ศพ	corpse	243				

From the subject word analysis, it can be inferred that the COVID-19 epidemic has been the main concern in Thailand's international news in recent years. The word "new coronavirus pneumonia" ranks first with a frequency of 1,429. In addition, there are other expressions of "new coronavirus pneumonia" with a total frequency of 2,292. The popularity is higher than all other events. In addition, keywords related to COVID-19 include "vaccine" with a frequency of 506 and an order of 16, "death" with a frequency of 409 and an order of 26, and "infection" with a frequency of 365 and an order of 26. 32. The frequency of "vaccine injection" is 254 and the order is 49. The frequency of "corpse" is 243 and the order is 52. The frequency of "sickness" is 187 and the order is 72. The frequency of "epidemic spread" is 160. The order is 96, the occurrence frequency of "patient" is 155, and the order is 99. Among the 30 content words before point h, there are 8 content words directly related to the epidemic, accounting for 26.67%. There are also words indirectly related to the epidemic, such as "blockade", and "control" and other words, the occurrence frequencies are 182 and 177 respectively, and the order is 77 and 83 respectively. It can be seen that the

epidemic situation is the top priority in international news in Thailand and is the main focus of the media. In addition, the frequency of occurrence of the word "military" is 216. After further analysis, 108 of the 216 news items involved are related to Myanmar. From this, it is not difficult to infer that Myanmar's military issue is another concern of Thailand. focus.

6. Conclusion

This article selects news materials from Thailand's mainstream media from August 2015 to August 2021, including Public Opinion, Frontline, Daily News, Thai Lat News, New News, Siam Politics, BBC (Thai), National newspaper and the Thai Post, a total of 650,039 news titles were obtained. Through classification tag screening, 8620 international news articles were obtained, accounting for 1.33%. After completing the collection of the news corpus, the author performed word segmentation by importing the Python library of PyThaiNLP. The Python library has For multiple Thai word segmentation algorithms, this article chooses the whitespace+newline algorithm. Through automatic word segmentation and manual review, a Thai news headline corpus was finally obtained with a sample size of 8620 headlines, a total of 6076 Types, and a total of 176,512 Tokens.

According to the h-point value calculated previously, which is 158.5, the author organized the top 99 high-frequency words as follows and translated the content words. According to statistics, there are a total of 97 words before point h, of which 30 are content words, accounting for 30.9%. Among these content words, there are 11 content words related to countries and regions, among which China, the United States, and Myanmar rank in the top three in terms of frequency. From the subject word analysis, it can be inferred that the COVID-19 epidemic has been the main concern in Thailand's international news in recent years. The word COVID-19 ranks first with a frequency of 1,429. In addition, there are other expressions of COVID-19, with a total frequency of 2,292. The popularity is higher than all other events. In addition, Myanmar's military issue is also another focus of international news attention in Thailand.

Funding

This paper is one of the stage results of the Philosophy and Social Sciences Planning Project of Guangdong Province, China, "Research on the Cross-Cultural Communication of China's National Image in Thailand under News Big Data" (Project Number: GD21YXW04).

References

- Chuanpit Tiemtanom. Analysis of the News Titles of "People's Daily" and "Xinhua Daily" (the meaning of words and the grammatical structure of phrases and sentences). *Journal of Chinese Studies*, Vol. 7 No. 2 (2014).
- Gu Dongxue. Analysis of news coverage of the "Belt and Road Initiative" in Thai English media from the perspective of evaluation theory [D]. Guangxi Normal University, 2019.
- Huang Tingting. Critical discourse analysis of Thai media reports on Chinese tourists [D]. Guangxi University for Nationalities, 2017.
- Kulnapa Siridissakul. China's national image in Thailand's mainstream media (2013-2019) [D]. Shanghai International Studies University, 2021. DOI:10.27316/d.cnki.gswyu.2021.001093.
- Le Yufei. Research on the characteristics of news headlines in Thai Rath newspapers [D]. Liaoning University, 2018.
- Nednamthip Buddawong. Political Discourse on COVID-19 Vaccine in Chinese Headlines, *Ubon Ratchathani University Art Journal*, June 30, 2022.
- Teewara Saengin. The language of online news headlines in mass media. *NBTC Journal*, 2020, 4(4), 374-393.
- Zhang Junjie. Research on linguistic characteristics in Thai media news headlines in multiple contexts [D]. Yunnan Normal University, 2021. DOI:10.27459/d.cnki.gynfc.2021.001071.