Canadian Journal of Language and Literature Studies
Vol. 2(5), 2022, pp.1-13

A Diachronic Shift in Japanese Word Length Distribution

Wenchao Li[1]

[1]Zhejiang University, China
Email: widelia@zju.edu.cn

DOI: 10.53103/cjlls.v2i5.64

Abstract

Given the typological differences between the Indo-European languages, which are fusional, and Japanese, which is agglutinative, the debate around the measuring unit of Japanese word length is unsurprising. This study delved into diachronic issues and calculated word length in Old, Early Middle, Middle, Early Modern, and Modern Japanese using data from eight writing systems, including 21 genres. This study aimed to clarify how word length distribution has shifted throughout history. The findings revealed that word length is associated with the writing system. Old Japanese bore the longest length, as it was utterly logographic. Since Early Middle Japanese, Japanese text has been written using a phonographic and logographic mix, and word length appears shorter. Furthermore, word length is associated with the diversity of genres. Moreover, an investigation of word length and frequency indicated that textbooks, sharehon, ninjoohon, and tales, which appeared after the Nara Period and used mixed Chinese character and kana writing, fit into the power law function.

Keywords: Japanese, Word Length, Genre, Measuring Unit, Mora, Syllable, Writing System

Introduction

Studies have confirmed that word length is associated with stylometrics (Mendenhall 1887, 1901; Williams 1975), frequency, with shorter words generally more frequently used (Zipf 1949), word age, ambiguity, and language acquisition proficiency at the early stage (Miyajima 1990; Sanada 1997; Ishii 1990; Ogino 1980; Minami et al., 2013). The present study delved into diachronic issues and examined eight writing systems, which included 27 genres from 11 historical periods, to explore how Japanese word length changed since the Nara Period (710–794) and whether the transitions of writing systems throughout history is associated with length distribution.

The investigation of Japanese word length is of particular interest in relation to quantitative studies for two reasons. When considering word length, two aspects are significant: the writing system and writing style (genre). The writing system is associated with language typology. Modern Japanese is phonologically moraic, morphologically agglutinative, and written using three scripts: Chinese characters, which are logographic,

and hiragana and katakana, which are phonographic. Summary (1) provides an illustration of the three scripts being combined.

(1) Hanako-wa-keeki-o-gatsugatsu-tabeteiru
Hanako-top-cake-acc-hungrily-eat-prog
'Hanako is eating cake hungrily.'

Japanese word length merits investigation, as the writing system has undergone a long-term evolution. During the Nara period, Old Japanese was used and the pure phonetic kana script had not been developed yet. Chinese characters were borrowed to represent vernacular on paper, leading to four scripts: Junsei-kanbun (purely classical Chinese), hentai-kanbun (variant Chinese), man'yōgana, and senmyō gaki. Summaries (2)–(4) provide illustrations of Old Japanese writing.

(2) ***Junsei-kanbun* 'purely classical Chinese'**

夜句茂	多菟	伊弩	毛夜覇餓岐	菟磨語味	爾
ya-kumo	tatu	idu	moya-pyegaki	tumagome	ni
many clouds	rise	Idumo	many-fenced palace	spouse dwell	DEM

夜覇餓枳	菟倶盧	廼	夜覇餓岐	廻
ya-pyegaki	tukuru	so no	ya-pyegaki	wo
many-fenced palace	build	DEM	many-fenced palace	ACC

(Nihon shoki kayō.1)

(3) ***hentai-kanbun* 'variant Chinese'**

曾能	那迦	都	迩	袁	加夫-**都久.**
so no	naka	tu	ni	wo	kabu-**tuku**
DEM	inside earth	DAT	ACC	wrap-attach	

'wrap the (three chestnuts) with earth.'

(KK.42)

(4) ***man'yōgana***

短	物	乎	端	**伎流**	等	云之….
Mijikaki	mono	o	hashi	***kiru***	to	ieru
Short.ACOP.ADN	thing	ACC	end	cut.CONCL	COMP	say.POTE

'We may say that we can cut one end of the short thing.'

(MYS.5)

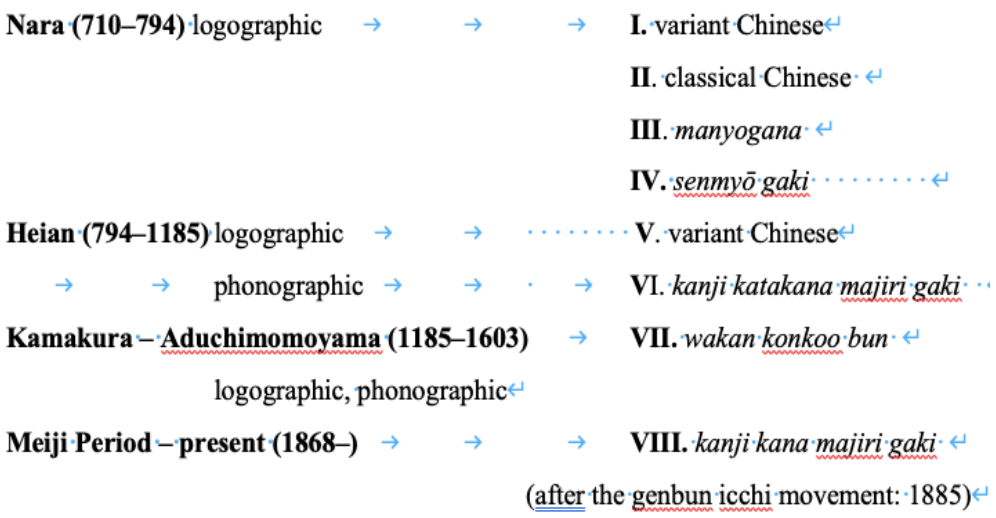
The categorisation of the development of the Japanese language is as follows.

a) Old Japanese (approximately 700–800 A.D.): Asuka and Nara Periods (710–794 A.D.).
b) Early Middle Japanese (800–1200 A.D.): Heian Period (794–1185 A.D.).
c) Middle Japanese (1200–1600 A.D.): Kamakura (1185–1333 A.D.), Muromachi (1336–1573 A.D.), and Aduchimomoyama Periods (1573–1603 A.D.).

d) Early Modern Japanese (1600–1868 A.D.): Edo Period (1603–1868 A.D.).
e) Modern Japanese (1868–Present): Meiji (1868–1912 A.D.), Taishoo (1912–1926 A.D.), Shoowa (1926–1989A.D.), Heisei (1989–2019 A.D.), and Reiwa Periods (2019–Present).

Man'yōgana paved the way for the pure phonetic script (kana), which was used in the Heian Period (794–1185 A.D.) and known as Early Middle Japanese. Consequently, the writing system consists of variant Chinese and kana. In the Late Heian period, mixed Chinese character and kana writing was developed. Together with variant Chinese, these scripts were used in writing such as poetry, diaries, essays, and setsuwa. Typical examples are Tosa Nikki (935), Taketori Monogatari (Early Heian Period), Konjaku Monogatari (1110–1124). During the Kamakura Period (1185–1333 A.D.), mixed Chinese character and kana writing continued, as represented by Heike Monogatari (before 1309). Two new scripts appeared: wakankonkobun (for essays) and sooroobun (for letters).

A summary of the shift of the Japanese writing system is as follows.



Accompanied writing system is the journey of text style. In Old Japanese, myths, poetry, chronicles, footprint poetry, and imperial poetry played a significant role. In Early Middle Japanese, diaries, tales, such as history and legends, essays, and setsuwa were significant. In Middle Japanese (Muromachi Period: 1336–1573), three new genres appeared: the linked verse, haikai, and Noh play. In the Aduchimomoyama Period (1573–1603 A.D.), the kyogen play was born. In Early Middle Japanese (Edo Period: 1603–1868), three styles were added: kana sooshi (essays written using mix of kana and Chinese characters), sharehon, and ninjoohon. In Modern Japanese (Meiji Period: 1868–1912), the genbun icchi movement, which was a unification of the spoken and written language,

encouraged writing in vernacular rather than classical Japanese. Meanwhile, foreign-origin words were borrowed into Japanese. Most of them were naturalized via three morphological devices: clipping (personal computer → pasokon), blending (yaburu 'break' + saku 'split' → yabuku 'break'), and acronyms (kokuritudaigakukyookai → kokudaikyoku 'the association of national universities'). With the genbun icchi movement and incorporation of foreign-origin words, modern Japanese word sources consist of Sino-Japanese, native Japanese, and loanwords, constituting a combination of logographic and phonographic writing. Therefore, a diachronic examination of the writing system and how it is related to word length distribution is required.

In this article, Section 2 outlines the methodology. Section 3 addresses word length distribution in the 11 periods and its sensitivity to the writing system. Section 4 discusses the associations between word length and genres. Section 5 presents the results and conclusion.

Methodology

Data

This study aimed to explore how Japanese word length has altered through time and illuminate the role of the writing system. To this end, a self-built database that included 11 historical periods, eight writing systems (variant Chinese characters, classical Chinese characters, manyogana, senmyō gaki, more variant Chinese, mixed Chinese characters and kana, and wakankonko bun), and 27 genres was examined. The materials are summarised in Table 1.

Table 1: Materials

Historical period	Materials	Genres
Nara Period (710-794)	Kojiki 712	Myth
	Fudoki 713	Poetry
	Nihonshoki 720	Chronicle
	Bussokuseki poetry 753	poetry
	Man'yōshū 759	Poetry
	Shoku Nihongi Senmyō 797	Imperial edict
Heian Period (794-1185)	Ryoounshu 814	Chinese poetry
	Kokin Wakashu 905	Japanese poetry
	Tosanikki 934	Diary
	Taketori Monogatari[1]	Tales in the form of legend

[1] Taketori Monogatari and Ise Monogatari are completed in early Heian Period, but no specific time has been known.

	Ise Monogatari	Tales in the form of poems
	Eega Monogatari 1092	Historical tales
	Makuranosooshi 1001	Essay
	Uchigikishuu 1134	Setsuwa
	Ryoounshu 814	Chinese poetry
Kamakura Period (1185-1333)	Shin Kokin Wakashū 1210	Japanese poetry
	Uji Shūi Monogatari 1221	Setsuwa
	Heikei Monogatari 1309	Military story
	Tsurezuregusa 1331	Essay
	Kanesawa Sadaaki Shojyoo 1308-1326	Letter
Muromachi Period (1336-1573)	Minasesanginhyakuin 1488	Linked verse
	Shinsen'inutsukubashuu 1524	Haikai
	Fuushikaden 1443	Noh play
Aduchimomoyama Period (1573-1603 A.D.)	Ookuratoraakirabon 1642	Kyogen play
Edo Period (1603-1868)	Ukiyo Monogatari 1665	Kana sooshi
	Yuushihoogen 1770	Share hon
	Shunshokuumegoyomi 1832	Ninjoo hon
Meiji Period (1868-1912) **Taishoo Period (1912-1926)** **Shoowa Period (1926-1989)** **Heisei Period (1989-2019)** **Reiwa Period (2019-)**	Kokumin no Tomo 1887-1898	Magazine
	Junjooshoogaku Kokugo Dokuhon 1933-1948	Textbook
	BCCWJ	Written Japanese
	CSJ	Spoken Japanese

Old Japanese data were extracted from the University of Oxford and the National Institute for Japanese Language and Linguistics (NINJAL) Corpus of Old Japanese, which is a lemmatized, parsed, and comprehensively annotated digital corpus of all Japanese texts from the Old Japanese period. Data for Early Middle, Middle, Early Modern, and Modern Japanese (Meiji Period [1868–1912 A.D.), and Taishoo Period [1912–1926 A.D.]) were collected from the Japanese historical corpus provided by the NINJAL. Written data for Modern Japanese (Shoowa [1926–1989A.D.], Heisei [1989–2019 A.D.], and Reiwa Period [2019–Present]) were obtained from the Balanced Corpus of Contemporary Written Japanese, whereas the spoken data for Modern Japanese were extracted from the Corpus of Spontaneous Japanese, both of which were produced by the NINJAL.

Word Length Measuring Unit

This study followed Popescu et al. (2013, p. 225) and used syllable numbers to evaluate Japanese word length in each period to ensure generality. A computer programme was created for calculations and fittings. The following procedures were carried out.

Step 1: Obtain raw data from the self-built database.

Step 2: Parse each sentence via the GiNZA v4 Parser (NINJAL and Megagon Labs).

Step 3: Romanise the Japanese scripts using a python programme.

Step 4: Calculate the dynamic mean word length distance from the parsed outputs based on syllable unit.

The associations between word length and genres were determined via Euclidean distance. Taking G1 and G2 as vectors representing the compared genres, the distance between L1 (G1,1, . . ., G1,n) and G2 (G2,1, . . ., G2,n) was calculated using the following formula:

Japanese Writing System and Word Length

Figure 1 shows Japanese word lengths through history, ranging from 1.71 to 2.534 syllables.

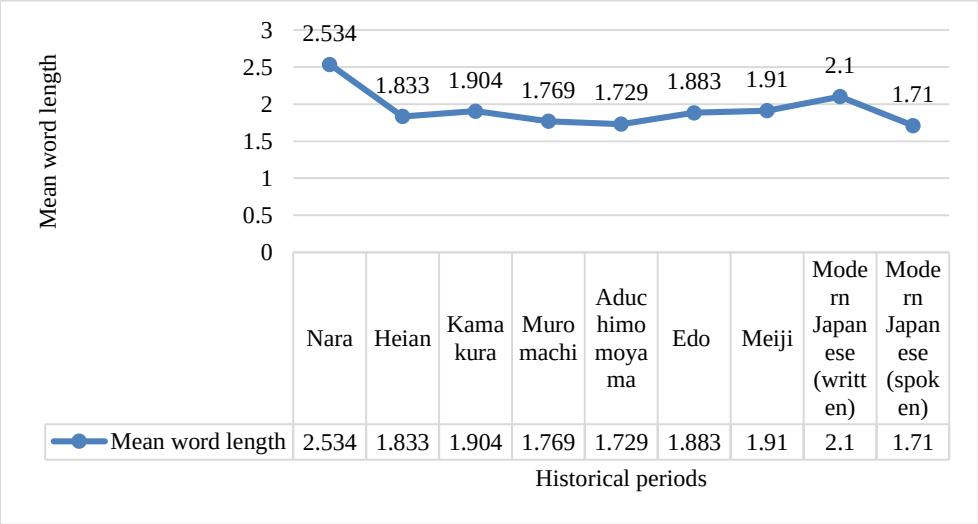
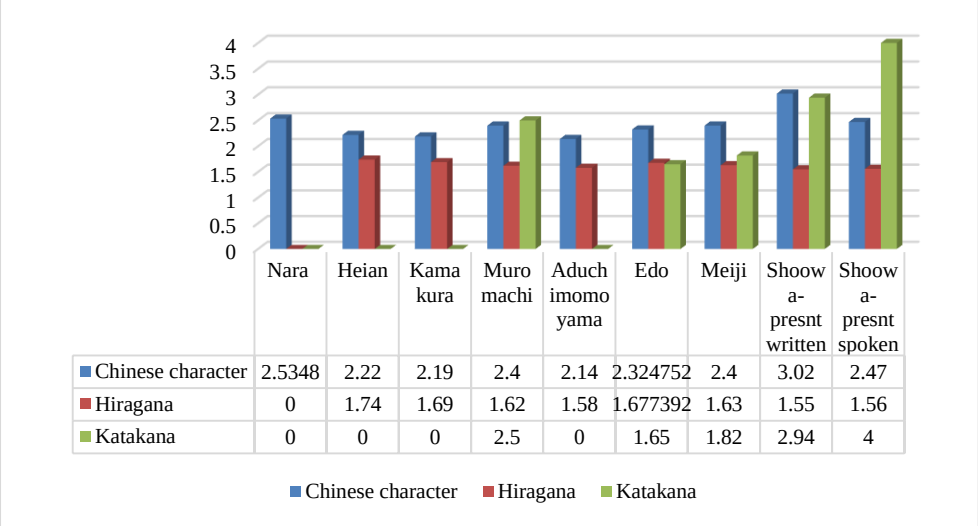


Figure 1: A shift of Japanese word length

Old Japanese was completely logographic. Its writings included variant Chinese, classical Chinese, manyogana, and senmyō gaki. Word length was the longest (2.534) in Old Japanese. The senmyogaki in Old Japanese was a mixed script of Chinese characters

and manyogana, specifically used for senmyo and norito. Nouns, adjectives, and verbs were written in big Chinese characters, and grammatical words, such as particles, auxiliaries, and suffixes, were writtern in smaller manyogana in the lower right. In the Heian Period (794–1185), manyogana developed into phonographic kana, and senmyogaki transformed into mixed Chinese character and kana writing. Kana refer to hiragana and katakana. Representative works include Tosanikki, Makuranosooshi, Genji Monogatari, and Taketori Monogatari. Mixed Chinese character and kana writing had a shorter word length (1.8) than variant Chinese (2.55), classical Chinese (2.96), and manyogana (2.8). In late Heian and early Kamakura Periods (1185–1333), Wakan konkobun, which is mixed Japanese and Chinese writing combining logographic and phonographic aspects, appeared. A representative work is the military story, Heike Monogatari. Middle Japanese established the Japanese writing system as a combination of phonographic and logographic writing. Essentially, the word length conveyed by phonographic and logographic appeared shorter than logographic scripts. In Modern Japanese, a split of length distribution is observed in spoken and written data.

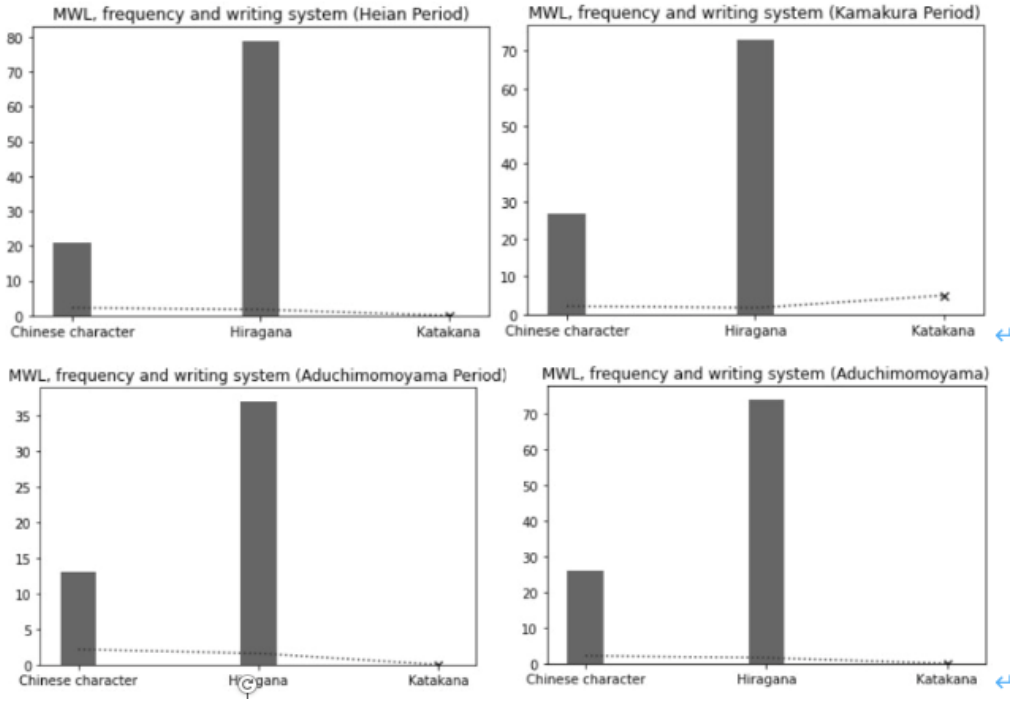
Another aspect of the Japanese lexicon is word sources. As touched upon earlier, during the Nara period (Old Japanese, AD. 710–794), before the development of phonetic script kana, Chinese characters were borrowed to represent vernacular Japanese on paper, which gave rise to four forms: Junsei-kanbun (purely classical Chinese), hentai-kanbun (variant Chinese), man'yōgana, and senmyō gaki. Man'yōgana turned into kana, a phonetic script that included hiragana and katakana, in the Heian Period. Hentai-kanbun (variant Chinese) was retained in the Heian Period and shifted into wakankonkoo gaki in the Kamakura Period. In Modern Japanese, loanwords were borrowed and naturalized into Japanese, written in katakana. Consequently, Modern Japanese writing consists of hiragana, katakana, and Chinese characters. We calculated the proportion of hiragana, Chinese characters, and katakana in each period and summarised the shifts in Figure 2.



	Nara	Heian	Kamakura	Muromachi	Aduchimomoyama	Edo	Meiji	Shoowa-presnt written	Shoowa-presnt spoken
Chinese character	2.5348	2.22	2.19	2.4	2.14	2.324752	2.4	3.02	2.47
Hiragana	0	1.74	1.69	1.62	1.58	1.677392	1.63	1.55	1.56
Katakana	0	0	0	2.5	0	1.65	1.82	2.94	4

Figure 2: A shift of word length and writing system

A more detailed depiction of the three writings and their word lengths and frequencies throughout the periods (except the Nara Period, as it was completely logographic) is provided in Figure 3.



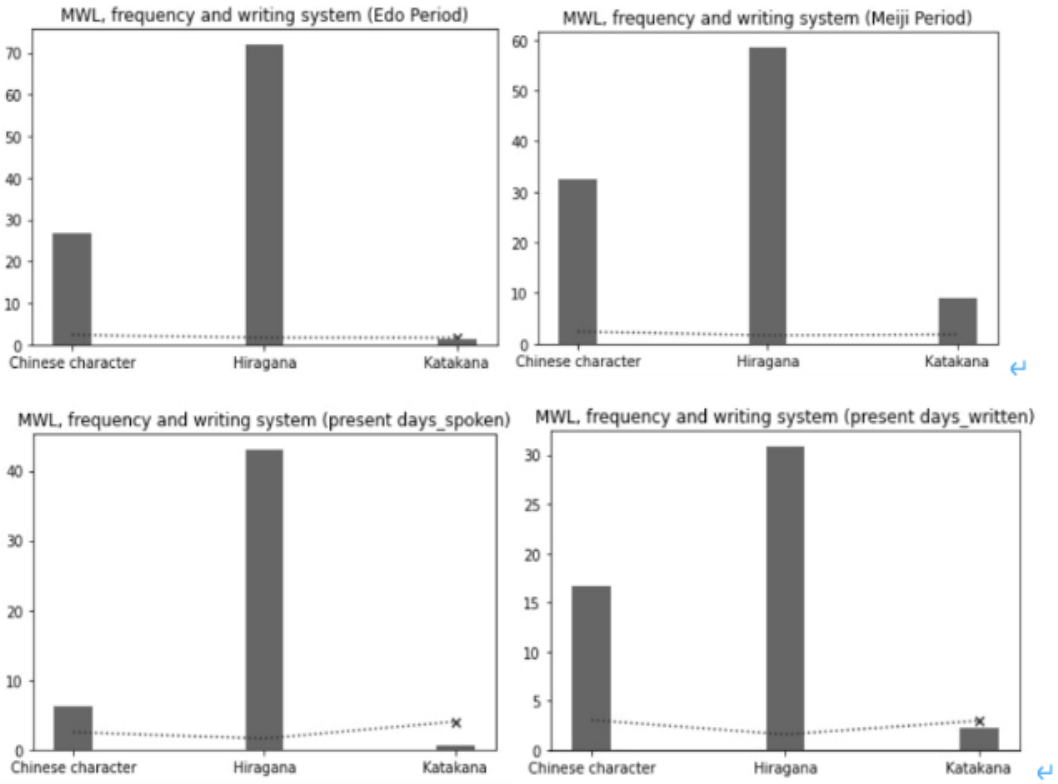


Figure 3: The three scripts and their word lengths and frequencies in each period

As shown in Figure 4, hiragana was mostly favoured since the Heian Period. Katakana originated in the Muromachi Period. A typical example is Shinseninutsukubashuu and Fuushikaden, where katakana frequency was 20% and the mean word length was 3. Shinseninutsukubashuu was a collection of Haikarenga from late Muromachi Period. Fuushikaden was written in mixed Chinese character and katakana script and addressed the theory and performance Noh technics. In the Edo period, katakana were detected in the sherehon yuushihoogen and had an average word length of 2. In the 'Kokumin's Tomo' magazine in the Meiji Period, katakana constituted 8.94% of the writing, with an average length of 1.82. The increasing use of katakana, was likely linked with the borrowing of foreign-origin words after the Meiji Restoration of 1868.

To summarise, the Japanese writing system has shifted from logographic script in the Nara Period (Old Japanese) to a combination of phonographic and logographic scripts since the Heian Period. Old Japanese had the longest word length. Lexicons written in the mix of phonographic and logographic scripts exhibited a shorter length. There was a split word length distribution in spoken and written data. Hiragana appeared to be mostly used among the three writing scripts in all historical periods, except Old Japanese.

Japanese Word Length and Genres

After establishing the associations between word length and writing system, we can explore the link between word length and genres. Texts of 21 genres were examined, including myths, poetry, chronicles, diaries, tales, essays, setsuwa, letters, verse, haikai, plays, kana sooshi, sharehon, ninjoohon, magazines, and textbooks. It should be noted that kyogen and Noh are spoken Japanese. Table 2 summarises the genres and their word length.

Table 2: Word length of various genres

Genres	Mean word length	Genres	Mean word length
Myth 712	2.55	Essay 1331	1.94
Poetry 713	2.55	Letter 1308-1326	2.13
Chronicle 720	2.96	Linked verse 1488	1.72
Footprint poetry 753	2.79	Haikai 1524	1.86
Poetry 759	2.8	Noh play 1443	1.72
Japanese poetry 905	1.7	Kyogen play 1642	1.72
Diary 934	1.77	Kana sooshi 1665	1.99
Tales in the form of legend (Early Heian Period)	1.83	Share hon 1770	1.81
Tales in the form of poems (Early Heian Period)	1.82	Ninjoo hon 1832	1.85
Historical tales 1092	1.92	Magazine 1898	2.01
Essay 1001	1.94	Textbook 1933-1948	1.82
Japanese poetry 1210	1.63	Written Japanese 1926-	2.1
Setsuwa 1221	1.88	Spoken Japanese 1926-	1.71
Military story 1309	1.94		

Euclidian clustering illustrated that myths, poetry (712), chronicles, and footprint poetry (753) had the longest word length. Letters, written Japanese (Modern), magazines, and kana sooshi fell into the middle-lengthened group. The short-length group included Japanese poetry, diaries, linked verse, Noh, kyogen, spoken Japanese (Modern), sharehon, tales (history and legends), textbooks, setsuwa, hakai, ninjoohon, essays, and military

stories. The categorisation is summarised in Summary (5).

(5)

Group A: long-length genre

myths, poetry (712), chronicles, footprint poetry (753)

Group B: middle-length genre

letters, written Japanese (Modern), magazines, kana sooshi

Group C: short-length genre

Japanese poetry, diaries, linked verse, Noh, kyogen, spoken Japanese (Modern), sharehon, tales (history and legend), textbooks, setsuwa, hakai, ninjoohon, essays, military stories.

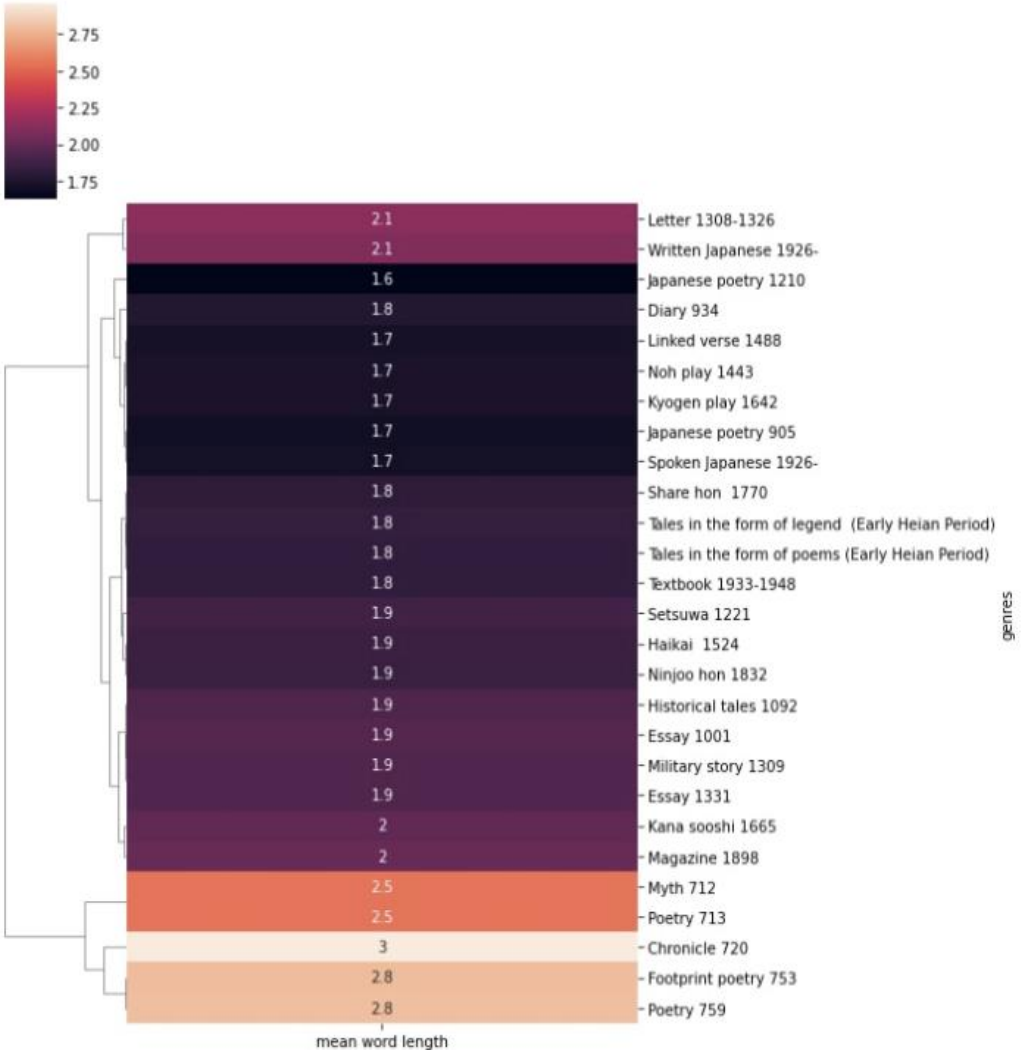


Figure 4: Genres and word length

Intriguingly, this categorization appears to be in accordance with the writing system. A further investigation of the word length and frequency in each genre was conducted. The finding indicated that textbooks, sharehon, ninjoohon, and tales (after the Nara Period and the onset of mixed Chinese character and kana writing) fit into the power law function. The fitting results are summarised in Figure 5.

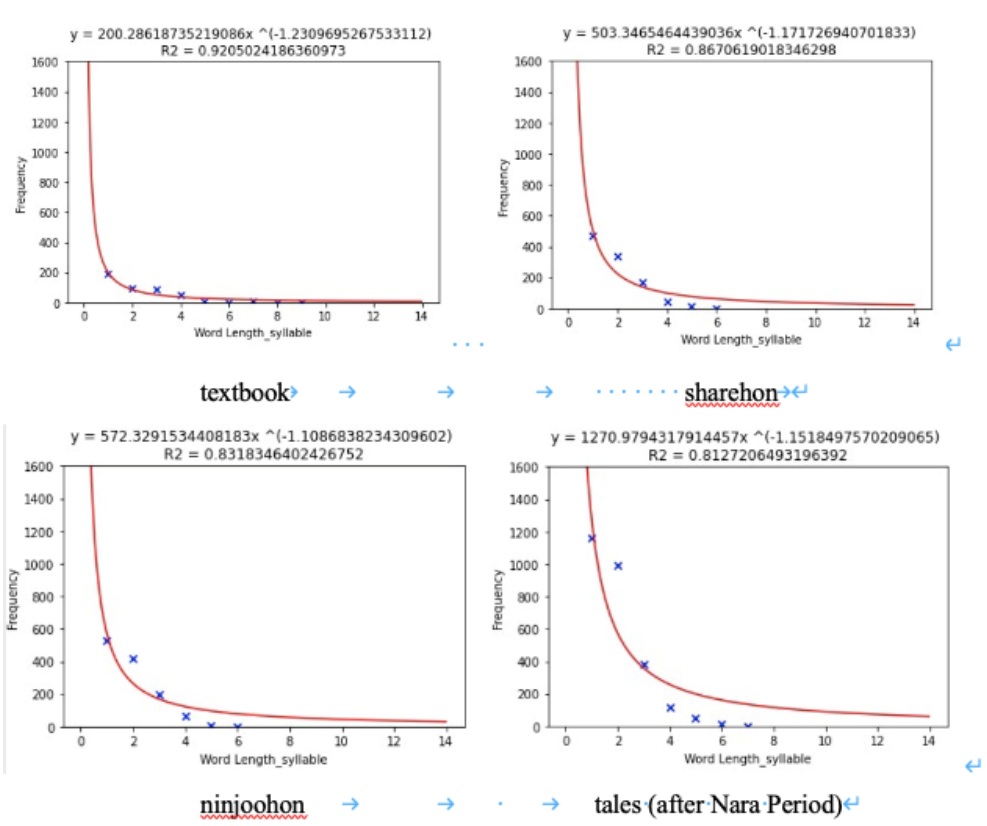


Figure 5: Fitting results of power law function to word length and frequency

Conclusion

This study examined Japanese word length in 11 historical period and explored changes through time, the role of the writing system, and the association between word length and genres. A self-built database, including eight writing systems (variant Chinese characters, classical Chinese characters, manyogana, senmyō gaki, more variant Chinese, mixed Chinese characters and kana, and wakankonko bun) and 27 genres, was examined.

The findings revealed that Japanese word length was associated with the writing system. Old Japanese was completely logographic and had the longest length. The Heian Period birthed phonographic kana. Combined logographic and phonographic writing had a

shorter length (1.8) than pure logographic writing in Old Japanese. Wakan konkobun, which is mixed Japanese and Chinese writing, in the late Heian and early Kamakura Periods paved the way for Modern Japanese writing. Furthermore, a split of length distribution was observed in spoken and written data.

Furthermore, Japanese word length was associated with genre. Euclidian clustering of texts of 21 genres categorised the genres into three groups: Group A included long-length writing, such as myths, poetry (712), chronicles, and footprint poetry (753), Group B included middle-length writing, such as letters, written Japanese (Modern), magazines, and kana sooshi, and Group C included short-length writing, such as Japanese poetry, diaries, linked verse, Noh, kyogen, spoken Japanese (Modern), sharehon, tales (history and legend), textbooks, setsuwa, hakai, ninjoohon, essays, and military stories. Moreover, textbooks, sharehon, ninjoohon, and tales (after the Nara Period and the onset of mixed Chinese character and kana writing) fit into the power law function.

References

Ishii, H. (1990). Word length in magazines [Zasshi ni okeru go no nagasa]. *Mathematic Linguistics [Keeryookokugogaku]*, 17 (4), 193-206.

Mendenhall, T. A. (1887). The characteristic curves of composition. *Science*, 11, 237-249.

Mendenhall, T. A. (1902). A mechanical solution to a literary problem. *Popular Science Monthly*, 60, 97-105.

Minami, Y., & Kobayashi, T. (2013). Correlations between word lengths and word acquisition times and periods of infants and toddlers. [Go no nagasa to yooji no goi shuutoku jiki, kikan to no sookan]. *Journal of the Phonetic Society [Onsee kenkyuu] 17*(3), 44-53.

Miyajima, T. (1990). Word frequency and length, age [Tango no shiyoo dosuu to nagasa, furusa]. *Mathematic Linguistics [Keeryookokugogaku] 17*(6), 287-300.

Ogino, T. (1980). Associations between honorification's length and degree of politeness: an investigation to Sapporo dialect (3) [heego hyoogen no nagasa to teeneesa: Sapporo ni okeru keego choosa kara]. *Mathematic Linguistics [Keeryookokugogaku] 12*(6), 264-271.

Sanada, H. (1997). The shift of Chinese translation in Meiji Period: A comparison of Philosophy lexicon and multiple vocabulary list [Meijiki kanyakugo no nagare: Tetsugaku jii to kakusyugoihyoo to no hikakuchoosa]. Presented at The 142nd Conference on Modern Japanese Language Research [Kindaigo kenkyuukai Dai 142 kai kenkyukai].

Williams, C. (1975). Word-length distribution in the works of Shakespeare and Bacon, Biometrika, 62, 207-212. Diederich.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge: Addison-Wesley.