9 Stylistic differences across time and register in Japanese texts: A quantitative analysis based on the NINJAL corpora

OGISO Toshinobu
National Institute for Japanese Language and Linguistics

Abstract
The construction of the Corpus of Historical Japanese (CHJ) is currently being prioritised at the National Institute for the Japanese Language and Linguistics. Thus far, the parts of the Corpus of Historical Japanese available to the public are the Heian period series and the Muromachi period series I: *Kyōgen*. The corpus of *Sharebon* (comprising data from a type of 18th to 19th century novel) and the *Kindai zasshi* corpus (comprising data from magazines of the 19th to 20th centuries) are currently under construction. These corpora are in the form of a full-text database, and are fully annotated with morphological information, such as parts of speech, lemma, and word form. Thus, multidirectional analysis of this data is possible.

As for historical Japanese documents, existing materials are limited, and we can usually use only the documents of a specific genre in a particular period. Therefore, it is often difficult to determine, for any particular characteristic discovered by observation of this corpus, whether it is the result of historical language change or due to a difference between genres.

In this paper, what will be demonstrated is an examination of the characteristics identified by the enumeration of the morphological information in the corpus of each historical period, and these will be compared to the characteristics of the various text genres of contemporary Japanese drawn from the Balanced Corpus of Contemporary Written Japanese. By these means we aim to investigate the characteristics of the documents constituting the CHJ more thoroughly and reliably, so that it may be used for historical study of the Japanese language.

Keywords: historical Japanese, corpus, language change, BCCWJ, CHJ

1 Background

When we study the history of the Japanese language, a fundamental problem is that the documents of each period are limited to particular genres. There are a few written documents from each historical period as most were scattered and lost over time. Furthermore, only a few of the available documents are suitable for the study of Japanese. For example, the most important material from the Nara period is limited to poetry from the *Man'yōshū* collection, and the most important documents for the study of language

in the Heian period are *kana* literary works such as *monogatari* novels and diaries. Although documents from non-literature genres have been preserved, most are written in the classical Chinese style and are not suitable for the study of Japanese. For periods after the Middle Ages, the situation is somewhat better and there are more remaining documents. However, documents that reflect spoken language are fewer in number. Thus, there are generally few historical documents suitable for thorough linguistic study.

In light of the above, we are forced to use documents of a specific genre in a particular period for studying Japanese. As a result, it is difficult to determine whether the characteristics observed in a historical corpus for any given period result from historical language change, or to differences among text genres. Mistaking stylistic differences based on text genre for difference due to time can be a serious problem for historical linguistic study. In this regard, it is important to understand the origin of any characteristic of the language in a corpus.

What will be proposed here is that the problem described above may be addressed by using data from both the Balanced Corpus of Contemporary Written Japanese (BCCWJ) and the current Corpus of Historical Japanese (CHJ). Differences in the Japanese language across genres may be identified from various types of texts in the BCCWJ, and differences across time may be identified from the historical texts of the CHJ. Thus, what will be shown here is a basic quantitative analysis of the data in various historical and contemporary Japanese texts.

2 Data source: The NINJAL corpora

At the Corpus Development Centre of the National Institute for Japanese Language and Linguistics (NINJAL), various types of corpora of Japanese have been and are being developed. They include data from Japanese language materials from ancient to contemporary Japan, both spoken and written. In this section, we will present an outline of the NINJAL corpora used for this study. Detailed information on these corpora appears on the website of the NINJAL Corpus Development Centre[1].

These corpora comprise full-text databases and are fully annotated with morphological information, such as parts of speech, lemma, and word form. The morphological information is based on the definition of Short Unit Words, defined by NINJAL for the purposes of the development of Japanese corpora (Ogura et al. 2011). The definition of the Short Unit Word is explained briefly in Section 2.4. By using morphological information, statistical analysis of text is possible.

1 http://pj.ninjal.ac.jp/corpus_center/en/

2.1 The Corpus of Historical Japanese

The CHJ is currently under development at the NINJAL. The aim of this is that the CHJ becomes a large diachronic corpus that covers historical Japanese language materials from the Nara period to modern times. As parts of the CHJ, the Heian period series and the Muromachi period series I: *Kyōgen* have been released.

The Heian period series contains fourteen *kana* literature works, namely *Tosa Nikki* (Tosa diary), *Taketori Monogatari* (The Tale of the Bamboo-Cutter), *Ise Monogatari* (The Tales of Ise), *Ochikubo Monogatari* (The Tale of Ochikubo), *Yamato Monogatari* (The Tales of Yamato), *Makura no Sōshi* (The Pillow Book), *Genji Monogatari* (The Tale of Genji), *Murasaki Shikibu Nikki* (The Diary of Lady Murasaki), *Izumi shikibu Nikki* (The Diary of Izumi Shikibu), *Heichū Monogatari* (The Tales of Heichū), *Sarashina Nikki* (Sarashina diary), and *Sanukinosuke Nikki* (The Diary of Sanukinosuke). These works were written in the Heian period (794–1185) and play a key role as Japanese classics, as they are of great literary value. Furthermore, these works have been widely used as the most important source of data for the study of Japanese in the Heian era.

The Muromachi period series I: *Kyōgen* contains 236 scripts of *Kyōgen* written by Okura Toraakira in 1642. *Kyōgen* is a form of comic theatre that developed alongside *noh* in the Muromachi period (1337–1573). For studying the language of the Muromachi era, *Kyōgen* scripts, Christian documents such as *Esopo no Fabulas*, and documents from *Shômono* (a kind of correspondence course) are important materials. Among these, *Kyōgen* plays a major role, and these texts have been used to study the spoken language of the Muromachi era.

In addition to this, a corpus of *Sharebon* is under development as part of the CHJ. This corpus is planned as part of the Edo period series of CHJ. *Sharebon* is a kind of novel developed in the 18th and 19th centuries. Twelve works of *Sharebon* are currently available. Although many documents written in the Edo era remain today, there are few that reflect the language that was spoken at that time. Therefore, *Sharebon* books, which contain many conversations, are valuable in the study of Japanese in the Edo era.

2.2 The *Kindai Zasshi* corpus

The *Kindai Zasshi* corpus consists of some independent corpora consisting of magazines published in the 19th and 20th centuries, earmarked for incorporation into the CHJ in the near future. One of the corpora is the *Taiyō* corpus, published in 2005, which is a corpus of the general interest magazine *Taiyō* (太陽, The Sun) published by *Hakubunkan* from 1895 to 1928. This magazine was read widely throughout Japan and had a widespread influence in the *Meiji* and *Taishō* eras. The *Taiyō* corpus contains the full-text of approximately 14,500,000 characters published in five years, namely 1895, 1901, 1909, 1917, and 1925. Because the *Taiyō* corpus includes a range of texts

by a large number of authors, it provides suitable material for the study of Japanese at the time. Although the *Taiyō* corpus was originally published without morphological information, morphological annotation has been completed, and morphological information has been manually corrected in the part that forms the core data. The core data is approximately 2% of the corpus.

The *Meiroku Zasshi* corpus, published in 2012, is another part of the *Kindai Zasshi* corpus. Although its scale is not large, it contains all issues of *Meiroku Zasshi* (明六雑誌, *Meiroku* Magazine) published by *Meiroku-sha* since 1874. It was the first modern magazine in Japanese that played a significant role in spreading Western ideas and thought. Its influence on Japanese cultural history was substantial, and it forms important material for the study of Japanese in modern times. The corpus comprises approximately 180,000 word tokens and its morphological annotation has been completely manually corrected.

Finally, the *Kokumin no Tomo* corpus, published in 2014, forms part of the *Kindai Zasshi* corpus. This is a corpus of *Tokumin no Tomo* (国民之友, The Nation's Friend) published by *Min'yū-sha* since 1887. This magazine corpus falls between the *Taiyō* and the *Meiroku Zasshi* period corpora. It was a widely read general interest magazine at the time. This corpus contains 101 million word tokens and a part has been manually corrected as the core data.

For the present study, we used the core data of the *Taiyō* corpus (222 thousand word tokens), all the data of the *Meiroku Zasshi* corpus (34 thousand tokens), and the core data of the *Kokumin no tomo* corpus (180 thousand tokens).

2.3 Balanced Corpus of Contemporary Written Japanese

The BCCWJ is a large-scale corpus of Japanese, containing more than 1,000 million word tokens. It includes contemporary written Japanese texts of various genres, media, and registers (Maekawa et al. 2014). The BCCWJ consists of both core data and non-core data. The morphological annotation of the core data has been manually corrected and its accuracy is higher than 99%. In contrast, the non-core data of the BCCWJ has been digitally analysed, but its accuracy is about 98%.

The BCCWJ is far larger than the historical corpora mentioned above. Therefore, the core data alone were considered sufficient for this study. The core data consist of six registers, namely books (PB: Publication+Books), magazines (PM: Publication+Magazines), newspapers (PN: Publication+Newspapers), white papers (OW: Out-of-population+Whitepapers), web data of the Q&A service "Yahoo! Chiebukuro" (OC: Out-of-population+*Chiebukuro*), and blogs (OY: Out-of-population+Yahoo! Blog). In this context, "out-of-population" means that texts of these sub-corpora were not sampled from a designed statistical population of contemporary written Japanese, but were collected for certain specific purposes.

2.4 The Corpus of Spontaneous Japanese

The Corpus of Spontaneous Japanese (CSJ), published in 2004, is a collection of large quantities of audio recordings of Japanese speakers, and is annotated with information for phonetic and phonemic studies. The CSJ also includes approximately 661 hours and 7,520 thousand transcribed tokens Although these data consist mainly of monologues, such as lectures, they serve as valuable contemporary spoken Japanese data.

In this study, we used approximately one million words from the text of the CSJ, called core data, as a sample of contemporary spoken Japanese (1,011 thousand tokens). The morphological information of the core data has been manually corrected.

2.5 Sizes of the corpora

Table 1 shows the sizes of the corpora described above.

Table 1 Sizes of the corpora

Target	Corpus	Sub-corpus	Size (tokens)
Diachronic (historical)	CHJ	Heian	871,477
		Kyōgen	277,424
		Sharebon	94,125
	Kindai	Meiroku	180,654
		Kokumin	34,279
		Taiyō	222,479
Synchronic (contemporary)	BCCWJ	OC_core	110,071
		OW_core	227,766
		OY_core	116,806
		PB_core	234,206
		PM_core	238,857
		PN_core	360,182
	CSJ	CSJ_core	1,011,681

Note: The names of corpora are abbreviated as follows:
Heian: CHJ, Heian period series (14 works)
Kyōgen: CHJ, Muromachi period series I
Sharebon: CHJ, Edo period series (12 works, under development)
Meiroku: Meiroku Zasshi corpus
Kokumin: Core data of the *Kokumin no Tomo* corpus
Taiyō: Core data of the *Taiyō* corpus

OC_core: BCCWJ core data of Yahoo! *Chiebukuro* data (Q&A web service)
OW_core: BCCWJ core data of government White Papers
OY_core: BCCWJ core data of Yahoo! Blog service
PB_core: BCCWJ core data of published Books
PM_core: BCCWJ core data of published Magazines
PN_core: BCCWJ core data of published Newspapers
CSJ_core: core data of the Corpus of Spontaneous Japanese

Except the case of *Sharebon* (now under development), the morphological information of all these sub-corpora has been manually corrected. In order to prioritise quality over quantity, the data for the present study were limited to manually corrected parts of the corpora, decreasing their size.

2.6 Historical periods and textual styles of the corpora

The table below shows the historical periods of the NINJAL corpora used for this study. Data were drawn from six historical and seven contemporary corpora.

Table 2 Historical periods of the corpora

Corpus	**Sub-corpus**	**Periods**	**Historical stage of Japanese Language**
CHJ	Heian Series	10-11c	Early Middle Japanese (*Chūko-go*)
	Kyōgen	15-16c	Late Middle Japanese (*Chūsei-go*)
	Sharebon	18-19c	Early Modern Japanese (*Kinsei-go*)
Kindai	Meiroku	1874-1875	Modern Japanese (*Kindai-go*)
	Kokumin	1887-1888	
	Taiyō	1895-1925	
BCCWJ	OC_core	2005	Contemporary Japanese (*Gendai-go*)
	OW_core	1976-2005	
	OY_core	2008	
	PB_core	1971-2005	
	PM_core	2001-2005	
	PN_core	2001-2005	
CSJ	CSJ_core	1999-2005	

The sub-corpora of the CHJ contain conversation-rich literary work, and are regarded as reflecting the colloquial language of the time. On the other hand, the sub-corpora of the *Kindai Zasshi* corpus contains a more refined text of the editorial writing style at the time.

The seven sub-corpora of contemporary Japanese contain a wide range of styles. OC and OY texts originate from web-based media, and are written in a light colloquial style. Conversely, OW texts are written in the formal style required for government white papers.

The terms "genre", "style", and "register" may be confusing and difficult to define. In this paper, the term "genre" is used to indicate the traditional classification of the document, such as a poem, novel, diary, playbook, editorial article, etc. The word "style" is used to refer to characteristics of the text, which derive from the differences among genres, entailing aspects such as historical era, tone of writing, etc. Finally, the word "register" is used for the various contemporary corpora and sub-corpora examined here. In the BCCWJ, sub-corpora such as OC, OW, PB, and PN are regarded as registers, and the CSJ is also regarded as a register of contemporary Japanese. The word register is purposely not used to refer to the historical corpora here.

2.7 Definition of word segmentation in the NINJAL corpora

In the Japanese writing system, no open space occurs between words, and agreement on how to segment words is generally lacking. Thus, it is necessary to define the manner in which words are segmented for annotation in the Japanese corpora.

The BCCWJ is annotated with morphological information in terms of units of two sizes, namely Short Unit Words (SUWs) and Long Unit Words (LUWs) (Maekawa et al. 2014). A SUW is a word of small size based on morphological unity. On the other hand, a LUW is a larger unit of words based on sentence structure (*bunsetsu*). A LUW consists of a combination of SUWs. For example, the string 国立国語研究所 (*Kokuritsu-kokugo-kenkyūsho*, National Institute for Japanese Language and Linguistics) is segmented as follows:

In SUWs: 国立 (*kokuritsu*, national) / 国語 (*kokugo*, Japanese language) / 研究 (*kenkyū* research) / 所 (*sho*, institute); four words

In LUWs: 国立国語研究所 (*Kokuritsu-kokugo-kenkyūsho*, literally: National-Japanese language-research institute); one word

The use of each of these two word units depends on the purpose of the study. However, except in the case of the BCCWJ and the *Heian* series of the CHJ, LUW annotation has not yet been completed. Thus, SUWs were used for the present investigation.

3 Measurement indices and results

In this section, we discuss the characteristics of the corpora on the basis of widely used indices, namely the parts of speech ratio, the type-token ratio (TTR), the modifier/verb ratio (MVR), and the *goshu* ratio.

Table 3. Numbers of tokens by parts of speech in the corpora

Corpus	Sub-corpus	Par-ticle	Verb	Copula	Aux. verb	Noun	Pro-noun	Adverb	Adj.	Adj. noun	Conj.	Prefix	Suffix	*Total*
CHJ	Heian	219058	163999	22831	73012	151792	12278	27127	35084	6394	521	14389	10874	*737359*
	Kyōgen	74818	47125	6665	20434	52999	8567	5220	5329	2193	1062	2512	2460	*229384*
	Sharebon	26279	13927	2310	5950	23416	2813	2948	2316	849	260	1402	1928	*84398*
Kindai	Meiroku	52274	28450	5225	10462	58422	4060	5824	2304	1456	2335	1063	2039	*173914*
	Kokumin	9624	4666	1077	2364	9189	679	1043	483	421	298	230	612	*30686*
	Taiyō	63616	29515	7132	12780	58238	4164	5049	3659	2468	1254	1330	4564	*193769*
BCCWJ	OC_core	30392	13404	2827	10378	26352	1425	1976	2182	1158	204	699	2101	*93098*
	OW_core	49157	21912	3537	5790	96297	433	811	1050	2510	1630	1715	11012	*195854*
	OY_core	27096	11335	2800	7891	31965	1375	2180	1851	1164	370	840	2592	*91459*
	PB_core	66616	30514	7520	14597	59821	3655	4208	3591	2675	871	1283	6064	*201415*
	PM_core	59876	25834	6276	11413	76812	2540	3568	3117	2613	553	1351	6446	*200399*
	PN_core	83719	33696	5594	14529	142680	1332	2154	3023	2882	602	2753	13992	*306956*
CSJ	CSJ_core	308060	129836	32002	87648	240674	21377	29414	14741	12729	11757	6079	20589	*914906*

Whereas the part of speech and type-token ratios are popular indices internationally, *goshu* is a unique index used for Japanese text analysis. *Goshu* refers to the origin of Japanese words (such as Chinese, Japanese, other foreign words, mixture of origins) and is similar to the strata of the English lexicon, which include words of Anglo-Saxon, French, Latin, etc. origin. The MVR is an index proposed by Kabashima (1965), and remains commonly used for Japanese text analysis (Koiso et al. 2008). The data relating to these indices were extracted from the NINJAL Morphological Information Database (Ogiso and Nakamura 2014) using SQL database queries.

3.1 Part of speech ratios

Table 3 shows the parts of speech ratios of the corpora. Although the corpora contain insubstantial tokens, such as signs, supplementary symbols, blanks, and particular un-analysable words, these are excluded from the data in the table. (A total of 469,000 tokens were excluded and approximately 96% of these were supplementary symbols and blanks.)

In all the corpora, the definitions of the parts of speech were based on the rules for SUWs. However, note that the auxiliary verbs *da* and *nari* were classified as verbs here. In Japanese, the behaviour of *da* (in colloquial language) and *nari* (in literary language) is similar to that of *be* verbs in English. They are therefore considered appropriate to be classified as copular verbs.

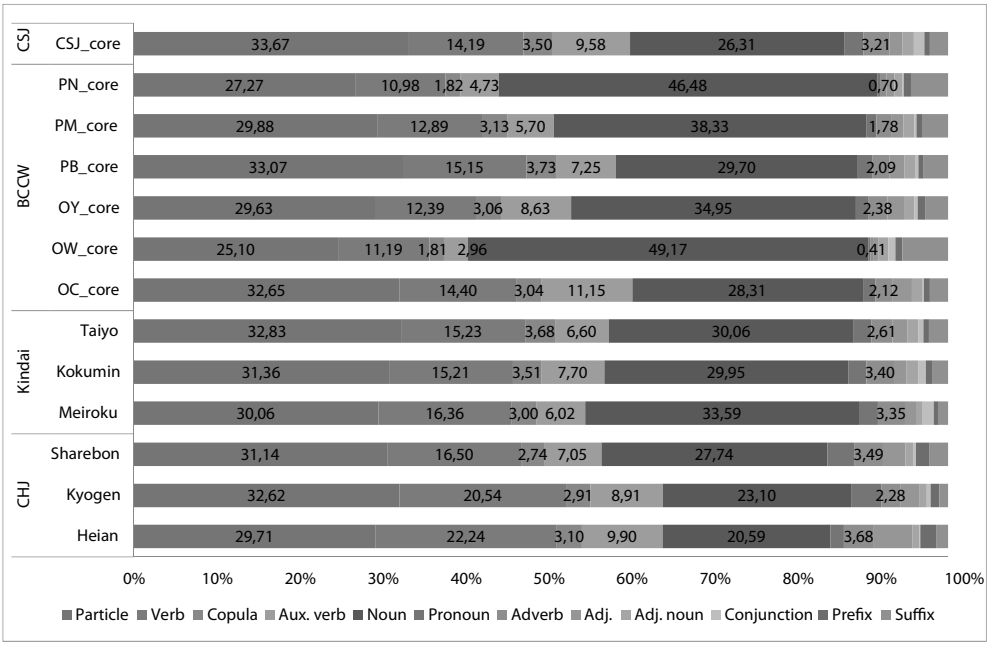

Figure 1: Parts of speech ratios based on tokens

Figure 1, which was generated with the data in Table 3, shows the parts of speech ratios of the corpora based on word tokens. For purposes of clarity, only the percentages of particles, verbs, copulas, auxiliary verbs, nouns, and adverbs are indicated.

With regard to parts of speech, it is known that the noun ratio reflects the characteristics of a text (Kabashima 1965, Koiso 2008). In the BCCWJ, sub-corpora written in in a refined style, such as OW and PN, have high noun ratios. However, in the corpora of *Kindai Zasshi,* also written in a formal style, the noun ratio is not as high. The high noun rates in OW and PN may be due to the multitude of compound nouns. In SUW analysis, compound nouns are divided into multiple nouns, resulting in an increase of the noun ratio.

From an historical perspective, the noun ratio increases over time: *Heian* series: 21%, *Kyōgen*: 23%, *Sharebon*: 28%, *Kindai*: 30-34%, and BCCWJ: 28-49%. Noun ratios including pronouns are as follows: *Heian*: 22%, *Kyōgen*: 27%, *Sharebon*: 31%, *Kindai*: 32-36%, and BCCWJ and CSJ: 29-49%. This finding may be due to historical changes, reflecting a consistent increase in the use of compound nouns. On the other hand, there remain considerable differences among the BCCWJ registers. Since a lot of nouns are divided into nouns and suffixes or prefixes and nouns in SUW annotation, this may influence the decrease in the number of nouns as well, which requires additional analysis based on LUW (in future).

3.2 Modifier/Verb Ratios

As mentioned above, the MVR is an index proposed by Kabashima (1965), and is calculated as follows:

$$\text{MVR} = \frac{\text{Number of modifiers (adjective + adjectival noun + adverb)}}{\text{Number of verbs}} \times 100$$

A high score means that the text contains many modifiers (words denoting manner), whereas a low score means that the text contains many verbs (words denoting movement).

The MVR scores of the corpora are as follows: *Heian*: 41.83, *Kyōgen*: 27.04, *Sharebon*: 43.89, *Kindai*: 34-42, and BCCWJ and CSJ: 20-45. These scores reflect straightforward historical change. The MVR scores are discussed in relation to the noun ratios in Section 0 below.

3.3 *Goshu* ratio

As mentioned above, *goshu* refers to the origin of Japanese words. Generally, Japanese words are classified in terms of three origins, namely *wago* (native Japanese words), *kango* (Chinese or Sino-Japanese words), and *gairaigo* (words of foreign origin other

than Chinese). Although most compound words consist of words with the same origin, some compound words consist of a mixture of *wago*, *kango*, and *gairaigo*. Such mixed origin words are labelled as *konshugo* (of hybrid origin). Such hybrid origin words are observed even in SUWs.

In the Japanese language, all particles and most adjectives (い形容詞) and auxiliary verbs are native Japanese, as are many basic words. On the other hand, Sino-Japanese and foreign words are common among nouns and adjectival nouns (形容動詞 / な形容詞). The origins of proper nouns are difficult to determine and are of less importance in the linguistic analysis of these texts. Therefore, they are simply marked as proper nouns without information about their origin.

Table 4 shows the numbers of *goshu* in the corpora. In this table, the "Other" column reflects errors of morphological analysis (i.e. unknown words) and words not defined for various reasons.

Table 4: Numbers of *goshu* in the corpora

Corpus	**Sub-corpus**	**Foreign (*gairaigo*)**	**Sino-Japanese (*kango*)**	**Hybrid (*konshugo*)**	**Native (*wago*)**	**Proper noun**	**Sign**	**Other**
CHJ	*Heian*	151	20696	2541	709814	4911	133309	61
	Kyōgen	207	18746	6197	207264	2434	42365	249
	Sharebon	149	7257	1668	72925	4606	7032	747
Kindai	*Meiroku*	558	43976	3677	127716	2655	2060	55
	Kokumin	59	6971	383	23256	643	2922	55
	Taiyo	617	50947	4060	150972	3298	4982	265
BCCWJ	OC_core	3423	16169	884	71717	1226	16761	102
	OW_core	4610	92214	2649	93712	2786	32118	83
	OY_core	4219	19135	1041	65281	2584	24782	200
	PB_core	4007	37755	1832	155227	5096	30431	53
	PM_core	10493	48367	2275	132597	7655	37999	60
	PN_core	9590	110045	2991	168325	16631	52865	79
CSJ	CSJ_core	24137	164910	7973	788933	10302	6027	13294

Figure 2 shows the *goshu* ratios in the corpora graphically. For purposes of simplicity, only the four main classes of *goshu* (*wago, kango, gairaigo*, and *konshugo*) are shown.

The ratio of words of native Japanese origin decreases over time as follows: *Heian*: 81%, *Kyōgen*: 75%, *Sharebon* 77%, *Kindai*: 68-71%, and BCCWJ and CSJ: 41-78%. On the other hand, the ratio of words of Sino-Japanese origin increases by the end of

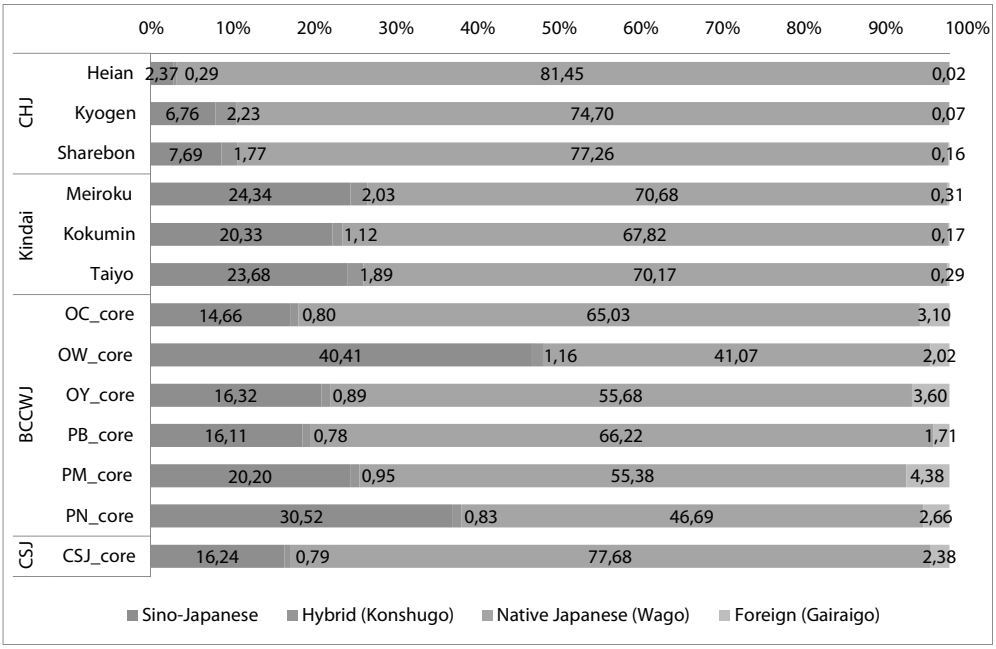

Figure 2: *Goshu* ratios in the corpora based on tokens (simplified)

modern period. However, the ratio of Sino-Japanese words may differ depending on registers in contemporary corpora. The presence of foreign words sharply increases in the contemporary corpora.

Thus, it appears that *goshu* may provide an indication of the historical period of a corpus. However, differences between registers remain considerable.

3.4 Type Token Ratio

The TTR is an index that explains the lexical density of a corpus. The value of the TTR is influenced by corpus size; thus, care is required in comparing TTRs across corpora. To deal with this problem, Baayen (2001) proposed indices that are more representative. However, for the present study, we simply took a designated number of words from the beginning of each corpus and calculated the TTR for that sample. As each text is automatically extracted from the beginning, the selected text represents a mixture of independent samples.

For example, the TTR indices for the Heian series were calculated as follows:

TTR (N=1000) = 299 (types) / 1000 (tokens)
TTR (N=10000) = 1235 (types) / 10000 (tokens)
TTR (N=100000) = 4192 (types) / 100000 (tokens)

Table 5 shows the results for the corpora studied here. The designated numbers of tokens for extraction were 1,000, 10,000, and 100,000, respectively.

Table 5: Type Token Ratios in the corpora

Corpus	Sub corpus	TTR		
		N=1,000	N=10,000	N=100,000
CHJ	*Heian*	29.90	13.25	4.19
	Kyōgen	29.30	11.76	4.56
	Sharebon	41.30	21.86	8.35
Kindai	*Meiroku*	38.00	19.41	9.59
	Kokumin	41.70	20.92	N/A
	Taiyō	39.30	20.69	12.97
BCCWJ	OC_core	33.30	18.57	8.12
	OW_core	27.20	13.43	4.86
	OY_core	36.40	22.08	9.85
	PB_core	32.20	15.85	8.38
	PM_core	31.30	18.25	9.83
	PN_core	31.90	21.00	10.31
CSJ	CSJ_core	28.50	10.37	3.30

The TTR scores in this table are difficult to evaluate in isolation. Similarly, small TTRs may be due to a number of different reasons in SUW analysis. For example, a small TTR may reflect the redundancy of spoken language. Thus, because data from historical literature (*Heian* and *kyōgen*) and the CSJ corpus are based on spoken language, their TTRs are small. However, if each text sample is long and includes many technical compound words and fixed phrase, the TTR will also be small. For example, the register OW leads to a small TTR because of the repetition of the same content words and similar expressions.

4 Analysis

Using the data presented in Section 3, we performed several statistical analyses, including an analysis of the relationship between the MVR and noun ratio, cluster analysis by *goshu* and parts of speech, and principal component analysis of synthetic indices.

4.1 Relationship between the MVR and noun ratio

In the field of the statistical analysis of Japanese texts, what has been accepted is that the relationship between the MVR and noun ratio (based on tokens) indicates characteristics of the text (Kabashima 1965). Kabashima states the following with regard to the MVR and noun ratio in contemporary Japanese:

1) If the ratio of the noun is big, and the MVR is small, it is an abstract text.
2) If the ratio of the noun is small, and MVR is big, it is a descriptive text.
3) If the ratio of the noun is small, and the MVR is small, it is a text tending to describe movement.

The above proposals were checked in terms of the historical data compared to the contemporary data in this study. Figure 3 shows a graph in which the MVR scores and noun ratios of the various corpora are mapped.

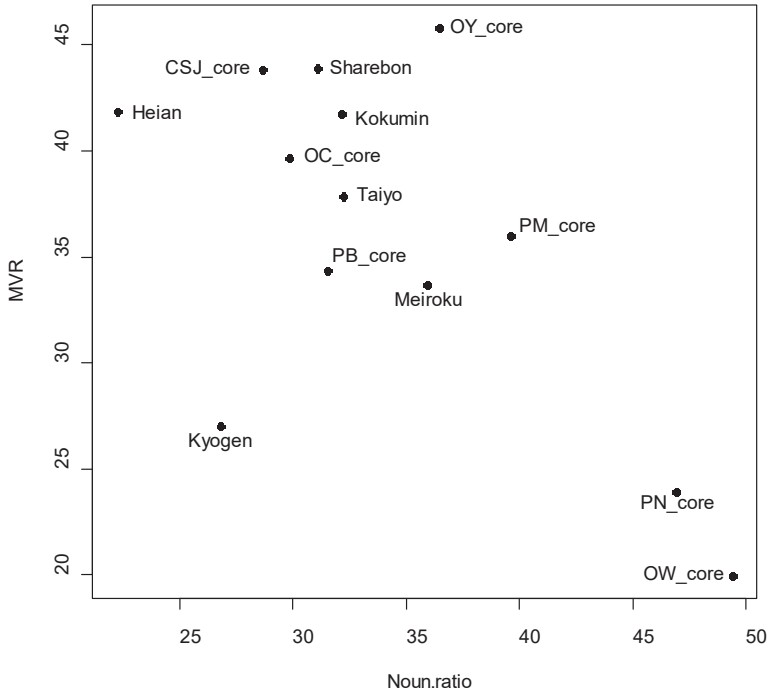

Figure 3. MVRs and noun ratios of the corpora

In Figure 3, the historical corpora and various contemporary corpora are mixed. In terms of Kabashima's (1965) proposal, OW and PN has a large noun ratio and

small MVR, which denotes them as most abstract texts in line with proposal 1. *Heian* is in line with proposal 2 as most descriptive data and CSJ is closest as well, as most descriptive and colloquial data. It is also understandable that *Kyōgen* describes movement.

However, there are some points more difficult to interpret. For example, the *Kokumin*, with the most refined editorial writing style, has been positioned near the colloquial *Sharebon* novels. Furthermore, although the noun ratios seem to explain, to a certain extent, the degree of formality or the casualness of the writing style, the *Kindai zasshi*, which contains mainly formal editorial texts, is located in the middle of the scale. The treatment of compound nouns in SUWs may have caused high scores in OW and PN, as pointed out in Section 3.1.

Thus, although Kabashima's (1965) explanation of the relationship between the MVR and noun ratio fits, to a certain extent, these corpora, it does not seem fully applicable to the present data, in which we analyse data across different periods based on SUWs.

4.2 Cluster analysis by *goshu* and part of speech

In this section, we will discuss the findings of cluster analysis using some of the indices discussed above, showing how the types of corpora are grouped together. As mentioned in Section 0, the *goshu* ratio reflects changes over time. The increase of both Sino-Japanese (*kango*) and Japanese native (*wago*) words are historically consistent. Furthermore, the differences across genres in contemporary Japanese are also considerable. However, words of foreign origin (*gairaigo*) suddenly increase in contemporary Japanese in comparison with *Kindai* corpora. It seems that the data may be grouped according to these features. To investigate this possibility, we performed cluster analysis using the *goshu* ratio. The cluster analysis was completed with the *hclust* function of R by using a Euclidean distance measure and Ward's methods.

Figure 4 shows the results of *goshu* cluster analysis. Only four main *goshu* features were used for this analysis, namely native, Sino-Japanese, foreign, and hybrid. The graph in Figure 3 shows that the corpora are clearly grouped into five clusters:

(1) Contemporary editorial writing style [OW and PN]
(2) Historical literary works [*Heian*, *Sharebon*, and *Kyōgen*]
(3) Contemporary text including conversations [CSJ, OC, and PB]
(4) 18th-19th century magazines (*Kindai Zasshi*) [*Kokumin*, *Meiroku*, and *Taiyō*]
(5) Other contemporary texts (contain no conversations, less bookish) [OY and PM].

Thus, it is confirmed that the *goshu* ratio is a useful index for classifying historical texts and those of various genres.

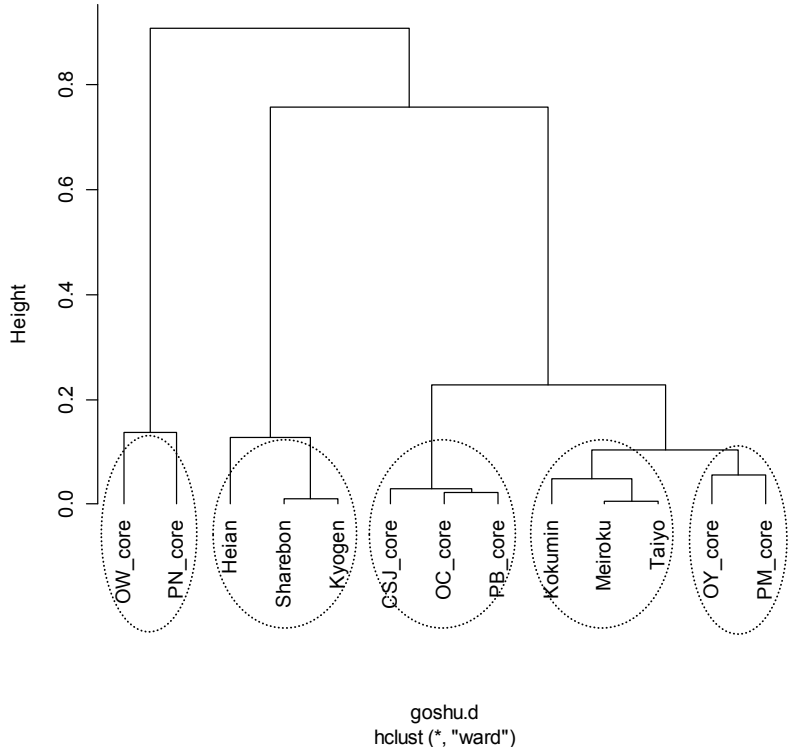

Figure 4. Cluster analysis by *goshu* ratio

Next, we performed a cluster analysis of the parts of speech ratios of the corpora, and compared the results to those of the *goshu* ratio. Figure 5 shows the results of the part of speech cluster analysis, which included 12 parts of speech, namely particle, verb, copula, auxiliary verb, noun, pronoun, adverb, adjective, adjectival noun, conjunction, prefix, and suffix. The analysis was performed in the same way as that for the *goshu* ratio above.

Whereas OW and PN, as well as OC and CSJ, are classified in the same way as in the *goshu* cluster analysis, the remaining corpora are clustered differently. In terms of characteristic writing style, the interpretation of this clustering seems difficult. Moreover, the clustering does not seem to reflect the expected characteristics across time periods.

As for the groupings by parts of speech data, it seems that various factors are overlapped, and classification in terms of stylistic differences cannot be made on the basis of parts of speech alone. As pointed out in Section 3.1, this may be due to the use of SUWs in the present study.

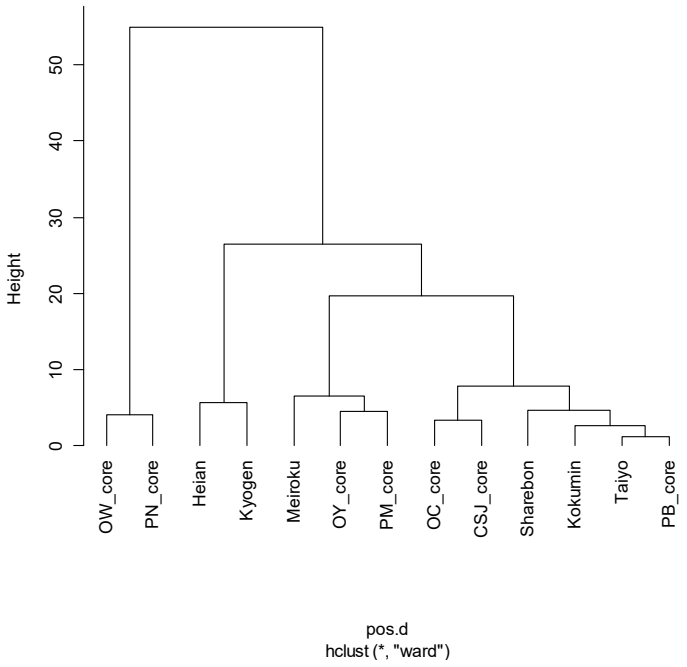

Figure 5. Cluster analysis by part of speech ratio

4.3 Principal component analysis using synthetic indices

Except in the case of the *goshu* ratio, individual indices were limited in their ability to explain the differences across the current text types. Therefore, we attempted to integrate the various indices discussed above, performing principal component analysis using general indices. Various combinations of characteristic features were analysed, including *goshu*, parts of speech, MVR, and TTR. Among them, the pair with the most explanatory power was *goshu* (*kango* and *gairaigo*) and parts of speech (nouns, modifiers, and verbs).

Figure 6 shows the results of the principal component analysis with the *goshu* and part of speech ratios. In this graph, the first principal component (PC1) appears on the X-axis and the second principal component (PC2) on the Y-axis. PC1 consists of noun: 0.4836604, modifier: -0.4735281, verb: -0.4753652, *kango*: 0.4585035, and *gairaigo*: 0.3250326. PC2 consists of noun: -0.2058790, modifier: 0.1899043, verb: -0.2150301, *kango*: -0.4068681, and *gairaigo*: 0.8424789.

The features of large absolute values contribute substantially to the axis. The arrows in the graph show the degree of the contributions to PC1 and PC2. In this analysis, the cumulative proportion of the principal components (PC1 + PC2) is higher than 0.934, and PC1 and PC2 explain most of the variation. Therefore, Figure 6 describes the data quite well.

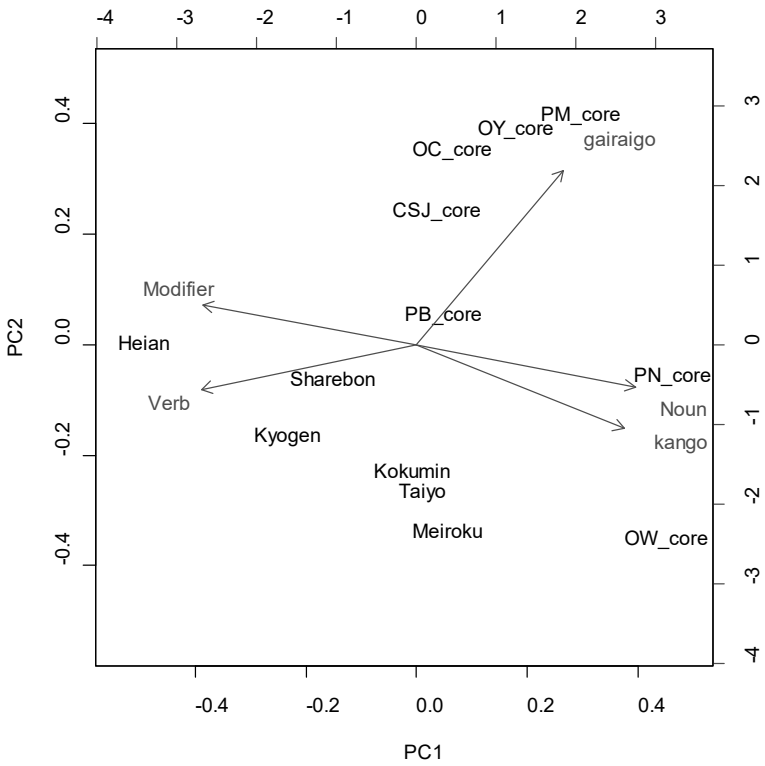

Figure 6. Results of principal component analysis

Note the almost exact chronological order of the corpora in relation to the PC1 axis: *Heian, Kyōgen, Sharebon, Kindai Zasshi*, and BCCWJ. Hence, PC1 may be taken as a component that describes historical differences. In addition, PC1 is also accurate in reflecting differences between colloquial texts (including conversation) and the editorial writing style in the BCCWJ registers. The PC1 axis leads to conversation-rich historical literary works. This axis is composed from parts of speech (+nouns, -modifiers and -verbs) and *goshu* (*kango* ratio).

The PC2 axis also distinguishes contemporary from historical texts. The percentage of *gairaigo* is the major component of PC2. Except for PN and OW, contemporary texts are located on the upper side and historical texts lie below. Moreover, for the BCCWJ registers, this axis shows the opposition between new and old in their lexicon.

In comparison with the wide range of the BCCWJ registers, the historical corpora are gathered together. Thus, the dispersion of the BCCWJ registers is greater than that of the historical corpora. As for the materials useful for the historical study of Japanese,

these tend to be conversation-rich in content, as colloquial language is regarded as important. As a result, whereas PC1 clearly showed differences across time, it also reflected the quantity of conversations in the respective corpora.

We also performed principal component analysis with the TTR. In this case, the cumulative proportion of the principal components (PC1 + PC2) was lower, around 0.82. The TTR has been rejected as the major variable composing PC2. Thus, the TTR seems to be independent from the parts of speech and *goshu* ratios.

5 Summary and conclusion

The present study showed that, for the classification of the mixed corpora of historical texts and various contemporary texts based on SUWs, the *goshu* ratio is the most effective index. The part of speech ratio alone was not enough for the classification of historical texts, as difference across genres greatly influenced the ratio. However, the parts of speech ratio may be more effective for an analysis based on LUWs, most historical texts are not yet annotated for this. Similarly, the MVR and TTR in isolation are not very effective indices. Classification based on both the *goshu* and parts of speech ratios was shown to be effective by principal component analysis.

Let us consider the position of Japanese historical documents in terms of these research results. The texts of the CHJ, examined here, are limited to *kana* literature works in the Heian period, script books of *Kyōgen*, *Sharebon* novels, and magazines from the 18th and 19th centuries. The results of this study showed these historical Japanese texts to be partial and one-sided in comparison with a variety of contemporary Japanese genres. Conversely, the dispersion of each register of contemporary Japanese is substantial.

In the study of historical Japanese, the content of documents is apt to be biased or partial. This must be borne in mind when we study the history of Japanese with corpora of such historical texts, and we should aim to include a variety of documents in the historical corpus after the Middle Ages, if such are available.

Finally, note that this research was based only on SUWs, which is not sufficient. We will strive to continue research with the entire vocabulary of the corpus, and to make these historical documents available for the study of the Japanese language.

Literature

Kabashima, T. and Jugaku A. (1965) *Buntai no kagaku*文体の科学 *(Science of the writing style).* Kyoto: Sogeisha. [In Japanese]

Baayen, R. H. (2001) Word frequency distributions, Kluwer Academic Publishers.

Koiso, H., Ogiso, T. and Ogura, H. (2008) Analysis of style in various genres based on *Short-Unit Word*, *Proceedings of the 2008 General Meeting of the MEXT Grant-in-aid for Scientific Research Priority Area Program "Japanese Corpus"*: 99-106. [In Japanese]

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M. and Den, Y. (2014) Balanced corpus of contemporary written Japanese, *Language Resources and Evaluation* 48(2):345-371.

Ogiso, T., Nakamura, T. (2014) Design, implementation, and operation of annotation support system for morphological information of BCCWJ, *Journal of Natural Language Processing* 20(1): 301-332. [In Japanese]

Ogura, H., Koiso, H., Fujiike, Y., Miyauchi, S., Konishi, H. and Hara, Y. (2011). Rule book of the morphological information for "Balanced Corpus of Contemporary Japanese" (『現代日本語書き言葉均衡コーパス』形態論情報規程集). Tokyo: National institute for Japanese language and linguistics. [In Japanese]

[Internet resources]

Center for Corpus Development, NINJAL: http://pj.ninjal.ac.jp/corpus_center/en/ (12.1.2016)

Corpus of Historical Japanese: http://pj.ninjal.ac.jp/corpus_center/chj/overview-en.html (12.1.2016)

要旨（Abstract in Japanese）

「各時代やジャンルにおける日本語文章のスタイル変遷–国立国語研究所の諸コーパスを利用した数量的研究—」

小木曽智信（国立国語研究所）

現在、国立国語研究所では『日本語歴史コーパス』の構築が進められている。これまでに「平安時代編」（10世紀から12世紀の仮名文学作品）、「室町時代編 I 狂言」（17世紀成立の狂言台本）のコーパスが公開されており、さらに、「洒落本」（18〜19世紀の小説の一種）、近代雑誌コーパス（19〜20世紀の雑誌）などのコーパスも構築されている。これらのコーパスは全文テキストを含むだけでなく、品詞や語彙素などの形態論情報が付与されているため、多角的な分析が可能になっている。

これらの古い時代の日本語資料は現存しているものが限られているため、一つの時代には特定のジャンルの資料しか利用できないことが多い。そのため、各時代のコーパスから得られる特徴が、その時代に特有のものなのか、そのジャンルに特有なものなのかを区別しづらい場合が少なくなかった。

そこで、本研究では、各時代のコーパスの形態論情報の集計結果から得られる特徴を、『現代日本語書き言葉均衡コーパス』から得られる現代語の多様なジャンルのテキストの特徴と比較して、時代的な差異と位相的な差異の両面から検討する。これにより、『日本語歴史コーパス』を構成するテキストの研究資料としての位置づけを考える。