WORD CLASS RATIOS AND GENRES IN WRITTEN JAPANESE: REVISITING THE MODIFIER VERB RATIO

Bor HODOŠČEK
Tokyo Institute of Technology,
Graduate School of Decision Science and Technology,
Department of Human System Science
hodoscek.b.aa@m.titech.ac.jp

Abstract

This paper explores the variability of genres in the Balanced Corpus of Contemporary Written Japanese using the modifier-verb ratio proposed by Kabashima and Jukaku (1965). Using bagplots to quantify the relation between noun and modifier-verb ratios for each genre, an attempt is made to classify genres on the scale of descriptive to summative according to Kabashima and Jugaku (1965). Our initial analysis confirms previous research results, while at the same time uncovering some contradictions in the ratios of the genre of magazines.

Keywords

BCCWJ, MVR, bagplot, genre, Long Unit Words

Izvleček

V članku bom raziskoval variabilnost žanrov v referenčnemu besedilnemu korpusu Balanced Corpus of Contemporary Written Japanese preko razmerja med dvema besednima vrstama – modifikatorjem in glagolom – prvič predstavljenega v delu Kabashime in Jugaku (1965) pod kratico MVR. Z uporabo statistične grafične metode bagplota raziščem relacijo med samostalniki in razmerjem med modifikatorji in glagoli, ter na podlagi te relacije klasificiram besedila v različnih žanrih po lestvici od opisnostni do povzetnostni. Analiza potrjuje večino prejšnih rezultatov, kot hkrati odkriva nekatera protislovja v relaciji med besednimi vrstami v žanru revij.

Ključne besede

BCCWJ, MVR, bagplot, žanri, dolge besedne enote

1. Introduction

The modifier-verb ratio, commonly abbreviated as MVR, was proposed in 1965 by Kabashima and Jugaku as part of a study on Japanese stylistics. More recently, as more sophisticated language processing tools and larger, more varied, corpora have become available, word class ratios have begun to be reexamined by studies like Fujiike, Konishi, Ogura, Ogiso, & Koiso (2011). In this paper, we will attempt to explore the variability between genres in the Balanced Corpus of Contemporary Written Japanese, as well as outline methodological issues in word class frequency extraction.

2. Materials and Methodology

This section introduces the corpus used for the proceeding analysis, including issues with sampling, balance and representativeness. Furthermore, the status of word units in a language without clear word boundaries and issues pertaining to word class classification and aggregate word class ratios like the modifier-verb ratio are explained.

2.1 The Balanced Corpus of Contemporary Written Japanese

Developed as part of the National Institute for Japanese Language (NINJAL) priority area research project "Japanese Corpus" [1], the "Balanced Corpus of Contemporary Written Japanese" (BCCWJ)[2] is a large scale 100 million word corpus constructed during the five year period 2006-2011. Aiming to represent contemporary standard written Japanese, it consists of written texts published in the 30 year interval from 1976 to 2005, except for several special-purpose corpora, which were by necessity collected only after the project had started. One aspect contributing to the balance of the corpus is the use of several different sampling methods that aim to faithfully represent the large body of contemporary written material in Japanese, as well as support different research goals and practical applications (Maekawa, 2007). Furthermore, the corpus consists of three sub-corpora, and the three sub-corpora consist of one or more genres (Table 1). One reason for including various specific purpose sub-corpora is because they consist of written material not in ordinary circulation, but without which the corpus could not be considered balanced. For example, with the advent of the Internet, a significant amount of language exchange happens under the radar, so to speak, of traditional media.

[1] Project homepage accessible at: http://www.ninjal.ac.jp/english/products/bccwj/.

[2] The 2009 project-internal (ryōikinai) edition is used here.

Table 1: BCCWJ sub-copora, genres, sampling period and LUW token counts used in this study

Sub-corpus	Genre	Sampling period	LUW tokens
Publication	Books	2001-2005	30,090,054
	Newspapers		1,018,274
	Magazines		2,363,513
Library	Books	1986-2005	25,509,426
Specific purpose	White papers	1976-2005	3,961,882
	Bestseller books		3,896,429
	Minutes of the Diet		4,659,941
	Textbooks	2005-2007	1,175,215
	Yahoo! Chiebukuro	2005	5,294,245
	Yahoo! Burogu	2008	2,723,005
			80,691,984

The Publication sub-corpus consists of randomly extracted samples from the population of all books, magazines, and newspapers in Japan. The Library sub-corpus is randomly sampled from the population of all books cataloged at more than 13 metropolitan libraries in Tokyo. The Specific purpose sub-corpus contains an unrelated set of corpora, ranging from government white papers, government-approved standard primary and secondary education textbooks, transcribed dialog from Japan's National Diet, and bestselling books, to blog and Q&A-style message board posts from Yahoo! Burogu and Yahoo! Chiebukuro, respectively. All three sub-corpora contain the genre of books, and as we are more concerned with characterizing inter-genre differences than intra-genre ones, we treat all book corpora as one corpus, thus lowering the total genre count to 8. In order to extract word class information from the corpora we introduce the concept of word units, natural language processing tools, and associated dictionary developed for the BCCWJ project.

2.2 Word units and UniDic

The concept of word units in Japanese is not as straightforward as in many Western languages, in large part because the Japanese writing system does not employ spaces to delimit word boundaries. One of the goals of the BCCWJ project was to provide a standardized unit of language that could accommodate diverse research goals and applications (Maekawa, 2007). The basic word unit taken up by the project was the Short Unit Word (SUW), which represents a relatively short unit corresponding to one morpheme. The second unit, called the Long Unit Word (LUW), is defined both in terms of SUW's and the phrasal bunsetsu unit in Japanese (Figure 1).

Bunsetsu	今回、		この	ホテルを		使って		大型夜景鑑賞イベントを					企画した。			
LUW	今回	、	この	ホテル	を	使っ	て	大型夜景鑑賞イベント				を	企画し		た	。
SUW	今回	、	この	ホテル	を	使っ	て	大型	夜景	鑑賞	イベント	を	企画	し	た	。
Reading	Konkai	,	kono	hoteru	o	tuka	tte	oogata	yakei	kansyo	ibento	o	kikaku	si	ta	.

Figure 1: Relationship between SUW's, LUW's and bunsetsu
(Source: Yomiuri shimbun (evening edition), 2004/4/28; BCCWJ sample ID: PN4c_00026)

SUW's are most simply defined as the dictionary entries contained in the morphological parser dictionary UniDic, which is a hierarchical dictionary specifically constructed for morphological parsing of the BCCWJ corpus (UniDic, 2010; NINJAL, 2011). The top-level word classes of the word class hierarchy contained in UniDic are nouns (名詞 /meisi/), pronouns (代名詞 /daimeisi/), verbs (動詞 /dousi/), i-adjectives (形容詞 /keiyōsi/), na-adjectives (形状詞 /keizyōsi/), adverbs (副詞 /hukusi/), prefixes (接頭辞 /settōzi/), suffixes (接尾辞 /setubizi/), interjections (感動詞 /kandōsi/), particles (助詞 /josi/), auxiliary verbs (助動詞 /zyodōsi/), pre-nominals (連体詞 /rentaisi/), conjunctions (接続詞 /setuzokusi/), symbols (記号 /kigō/), punctuation (補助記号 /hozyokigō/), and space characters (空白 /kūhaku/) (UniDic, 2010). As LUW's are constructed from SUW's, the word class hierarchy is identical for the most part. One important difference, especially for this study, is that ambiguous SUW's that, depending on context, can be either a noun or adverb, or a noun or verb, are disambiguated in the process of becoming LUW's. This enables the computation of more accurate modifier-verb ratios that are also more comparable to the ones calculated by Kabashima and Jugaku (1965).

2.3 Modifier-verb ratio

According to Kabashima and Jugaku, texts can be classified on a scale ranging from summative (要約的 /yōyakuteki/) to descriptive (描写的 /byōsyateki/) (1965). Summative texts convey only the bare minimum – the skeleton or framework of what they are describing. In contrast, descriptive texts specify in detail what they are describing, making the reader feel as if he is part of the situation described. For descriptive texts, Kabashima and Jugaku (1965) further characterize them as active (動き描写的 /ugokibyōsyateki/) or static (ありさま描写的 /arisamabyōsyateki/).

The modifier-verb ratio was first introduced in Kabashima and Jugaku (1965) as a quantitative method of classifying texts into these categories using only the ratio of modifier to verb counts. Kabashima and Jugaku define modifiers as consisting of adjectives, adverbs and pre-nominals (1965, p. 122). Having classified words into word

classes and calculated ratios based on each word class, one can then calculate the modifier-verb ratio by dividing the ratio of modifiers with the ratio of verbs in a text:

$$MVR = 100 \times \frac{modifiers}{verbs}.$$

In sum, texts with a high noun ratio and low modifier-verb ratio tend to be summative, while those with low noun ratios tend to be descriptive (see Figure 1). Furthermore, descriptive texts with low modifier-verb ratios tend to be active, while those with high modifier-verb ratios tend to be static.

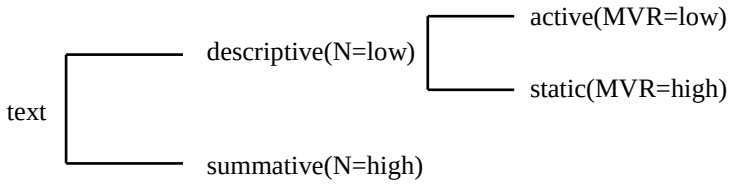

Figure 2: Categorization of texts using noun and modifier-verb ratios (adapted from Kabashima p. 25; N and MVR information added by author)

2.4 Extraction of modifier-verb ratios

Modifier-verb ratios for all samples in the BCCWJ are computed as follows. First, using plain text samples from the BCCWJ, we split sentences based on a set of common sentence delimiters (from the set of half- and full-width characters ". !?. 。 ！？"), except for when such a delimiter is encountered inside a quotation. We then morphologically analyze the sentences into SUW's using the morphological parser MeCab version 0.98 and UniDic version 2.1.0 (Kudo, 2010; UniDic, 2010; NINJAL, 2011). In the next step, the SUW data is fed into the LUW analyzer Comainu, version 0.53a, which produces morphologically parsed Japanese text with LUW's as the base unit. Finally, following Kabashima and Jugaku (1965), all word classes are coded and counted into the five classes of modifiers (M), nouns (N), verbs (V), interjections (I), and other (O). Using these frequency counts, it is straightforward to calculate the modifier-verb ratio as MVR = 100 x M/V. Special care has to be taken for samples that contain no verbs, and for these we do not calculate the modifier-verb ratio, but treat them as outliers (572 and 492 samples from Yahoo! Chiebukuro and Yahoo! Burogu, respectively).

Kabashima and Jugaku (1965)'s methodology differs slightly from the one used here, for they took random sentences as the sampling method from each book, giving them the average word class ratio of each book. In contrast, for at least the library and

publication sub-corpora in the BCCWJ, the sampling was done on a whole body of material, with samples taken in fixed- and variable-length chunks from random books and other media, and thus while possibly misrepresenting the sampled work, the samples, when taken together, offer a representative sample of the body sampled (Maruyama et al., 2010).

3. Results

Having defined modifier-verb ratios and explained their extraction procedure in the previous section, this section plots the relationship between noun and modifier-verb ratios using bagplots. In addition to visual information, several summary statistics based on bagplots are also provided and used to quantify genre based on their distributional properties.

3.1 Bagplots of noun and modifier-verb ratios

Compared to Fujiike et al. (2011), whose study was based on smaller sample sizes and used a scatterplot to visualize noun and modifier-verb ratios, we plot the relationship between noun and modifier-verb ratios for each genre using a bivariate visualization method called a bagplot. The bagplot, first proposed by Rousseeuw et al. (1999), is a 2D generalization of a boxplot used to analyze the relationship between two variables. According to Rousseeuw et al. (1999), "the bagplot visualizes the location, spread, correlation, skewness, and tails of the data." More specifically, it consists of a bag containing 50% of the data points, a fence (computed by magnifying the bag by a factor of 3) separating inliers from outliers, and a loop containing points outside the bag but inside the fence. In addition, the central point for each bagplot is defined as the point with the highest halfspace depth and is denoted by a star-shaped point, while any point outside the fence is considered an outlier.

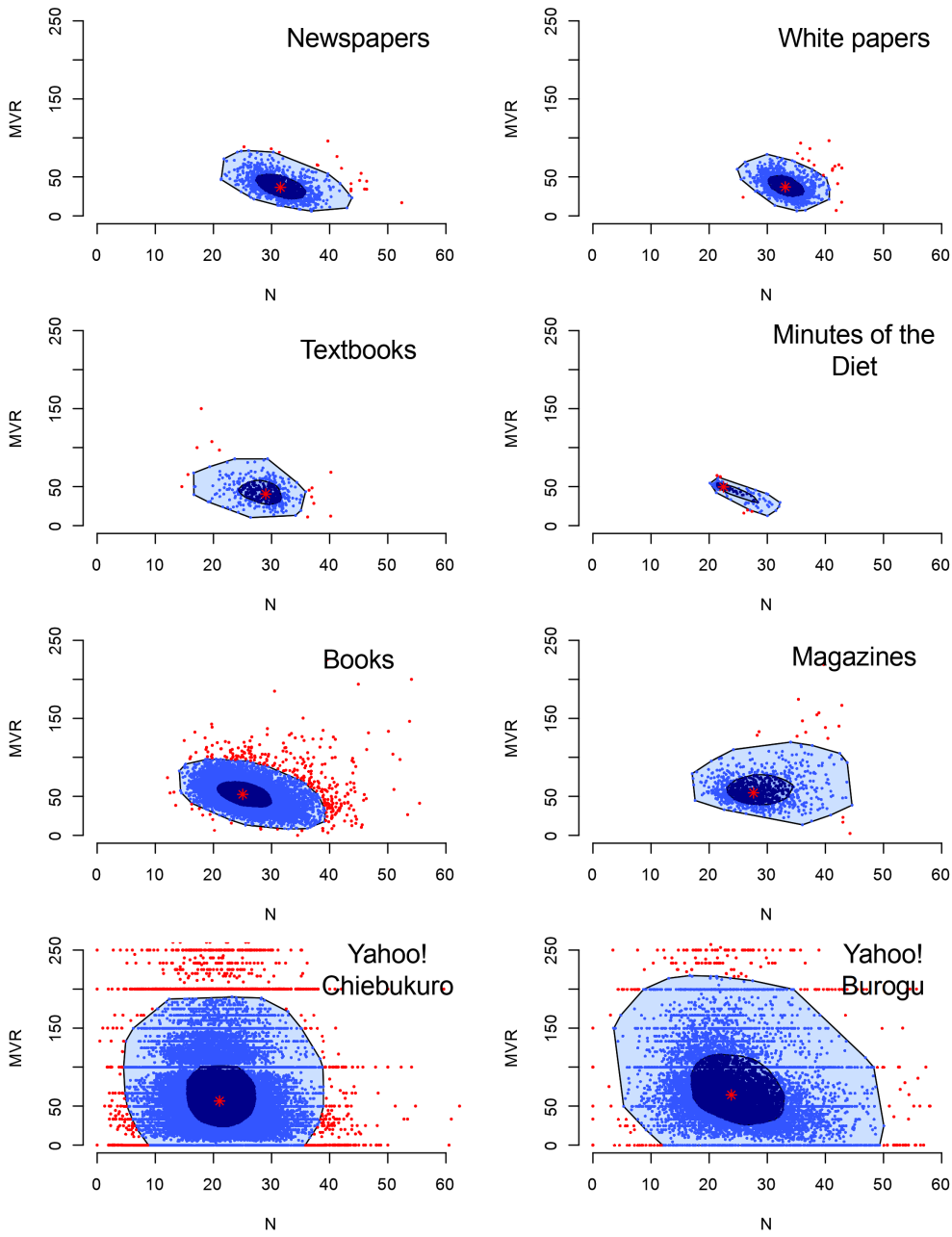

Figure 3: Bagplots of noun ratio (N) to modifier-verb ratio (MVR) for each genre.

Intuitively, the location of each genre is the halfspace median, here denoted by the central star-shaped point; the spread is represented by the size of the bag; the correlation between N and MVR by the orientation of the bag; skewness by the shape of the bag and loop; tail by points near the fence and beyond.

This study uses the ratios computed with the method outlined in subsection 2.4 together with the aplpack package for the statistical programming environment R to plot the bagplots (Wolf and Bielefeld, 2010; R Development Core Team, 2010). Figure 3 shows bagplots plotting the relationship between noun and modifier-verb ratios for each genre, here limited to noun ratios of 0-60% and modifier-verb ratios of 0-200 for clarity.

Table 2: Summary of bagplot statistics for each genre, sorted by MVR median

Genre	Bag (%)	Fence (%)	Outliers (%)	N median	MVR median
Newspapers	51,17	47,45	1,37	31,50	36,35
White papers	49,47	49,40	1,13	33,07	36,93
Textbooks	49,69	47,62	2,69	29,00	40,82
Minutes of the Diet	49,06	48,43	2,52	22,46	49,86
Books	52,36	46,22	1,41	25,00	52,86
Magazines	50,33	48,14	1,53	27,63	54,96
Yahoo! Chiebukuro	53,91	41,35	3,66	21,01	56,82
Yahoo! Burogu	48,51	43,28	3,71	23,76	64,77

Furthermore, in order to enable easier comparison and quantitative assessment of each genre, we also tabulate the halfspace median of N and MVR, the percent of samples inside the bag, the percent of samples outside the bag but inside the fence, as well as the percent of outliers for each genre (Table 2).

4. Discussion

In general, the negative correlation of noun ratios with modifier-verb ratios observed by Kabashima and Jugaku (1965) was re-confirmed for the BCCWJ in general, as can be seen from the general bottom-down facing orientation of the bags. The genre of Magazines was the only exception to this tendency, and merits further investigation. Interestingly, Magazines have both a relatively high noun ratio as well as a high modifier-verb ratio, a combination not adequately treated in Kabashima and Jugaku (1965).

Although both Internet corpora, Yahoo! Burogu and Yahoo! Chiebukuro, showed the biggest bag and fold areas, as well as the highest outlier percentages, Yahoo! Chiebukuro, in particular, has the highest concentration of samples in the bag and least between the fence and bag, suggesting a different distribution of word classes than other genres.

Not surprisingly, White papers, Newspapers, and Textbooks have the highest noun ratios, as well as the lowest modifier-verb ratios, classifying them as summative texts, while Books has an average noun ratio, but higher than average modifier-verb ratio.

5. Conclusion

This paper has hoped to shed some light on the ways in which the modifier-verb ratio can be applied to the study of genres. Issues in the extraction of word class ratios pertaining to word units and the morphological parser dictionary UniDic were touched on. Finally, using the bagplot to visualize the distributions of word classes inside genres and between genres was attempted, but revealed that further study was necessary to uncover the causes of variation and deviations from previous research in connection with the positive correlation observed for Magazines.

6. Future Work

The advent of the BCCWJ, its inclusion of various sampling targets and genres, in particular, lowers the barriers for the comparative study of genres in Japanese, as well as decreases the likelihood of overextending generalizations from one narrow genre, such as newspaper Japanese, onto the whole of Japanese. Modifier-verb ratios are a relatively simple index of variety observable in and between genres, and more needs to be done to extend Kabashima and Jugaku's analysis to new written genres. Additionally, comparisons with other measures like lexical density, as for example in Halliday (2009, p. 75-77), or the multidimensional feature approach outlined in Biber and Conrad (2009) should be attempted.

Another issue not adequately addressed here is the role of topic in genre studies. As a word class based measure, the modifier-verb ratio should be relatively robust to topic changes in an otherwise situationaly homogeneous genre. However, this should be qualified by, for example, using the Nippon Decimal Classification (NDC) library classification system supplied for the book corpora, as well as the topical information available for the Internet corpora to further understanding of intra-genre variation phenomena.

Finally, though the BCCWJ contains material sampled from a relatively short timespan of up to 30 years, we cannot be sure that any inference we make from the data is not attributable to diachronic differences.

References

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge Textbooks in Linguistics.

Fujiike, Y., Konishi, H., Ogura, H., Ogiso, T., & Koiso, H. (2011, Mar.). Tyōtan'i ni motozuku 'gendai nihongo kakikotoba kinkō kōpasu' no hinsihiritu ni kansuru bunseki. In *Proceedings of the 17th Annual Meeting of The Association for Natural Language Processing* (Vol. 17, pp. 663-666). Toyohashi, Japan.

Halliday, M. (2009). Methods - techniques - problems. In M. Halliday & J. J. Webster (Eds.), *Continuum Companion to Systemic Functional Linguistics* (pp. 59-86). Continuum International Publishing Group.

Kabashima, T., & Jugaku, A. (1965). *Buntai no kagaku* [Stylistics]. Sogeisha.

Kudo, T. (2011, July 20). MeCab: yet another part-of-speech and morphological analyzer. Retrieved from http://mecab.sourceforge.net/

Maekawa, K. (2007). Kotonoha and BCCWJ: development of a balanced corpus of contemporary written Japanese. In *Proceedings of the First International Conference on Korean Language, Literature, and Culture* (Vol. 2, pp. 158 −177). Corpora and Language Research. Seoul.

Maruyama, T., Yamazaki, M., Kashino, W., Sano, M., Akimoto, M., Inamasu, S., & Oyauchi, Y. (2010). Outline of sampling method in the balanced corpus of contemporary written japanese (4): Corpus design and the result of sampling. In *Tokutei ryōiki kenkyū 'nihongo kōpasu' heisei 21 nendo kōkai waakusyoppu (kenkyuseikahōkokukai) yokōsyū* (pp. 37-46).

NINJAL [National Institute for Japanese Language and Literature]. (2011). Tokuteiryōiki kenkyū nihongo kōpasu kenkyū seika hōkoku [Priority-Area Research "Japanese Corpus": Research Report] [DVD media containing UniDic and Comainu]. Tokyo: General Headquarters, Priority-Area Research "Japanese Corpus".

R Development Core Team. (2011, July 20). *R: a language and environment for statistical computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from R Foundation for Statistical Computing: http://www.R-project.org

Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician*, 53(4), 382-387. doi:10.2307/2686061

UniDic. (2011, July 20). Keitaiso kaiseki zisyo UniDic. Retrieved from http://www.tokuteicorpus.jp/dist/

Wolf, P., & Bielefeld, U. (2011, July 20). *Aplpack: another plot package: stem.leaf, bagplot, faces, spin3r, and some slider functions*. R package version 1.2.3. Retrieved from http://CRAN.R-project.org/package=aplpack