An Approach toward Register Classification of Book Samples in the Balanced Corpus of Contemporary Written Japanese

Wakako Kashino[ab] and Manabu Okumura[b]

[a]Department of Corpus Studies, National Institute for Japanese Language and Linguistics,
10-2 Midoricho, Tachikawa City, Tokyo, 190-8561, Japan
waka@ninjal.ac.jp

[b]Precision and Intelligence Laboratory, Tokyo Institute of Technology,
4259 Nagatsuta, Midori-ku, Yokohama, 226-8503, Japan
oku@ lr.pi.titech.ac.jp

Abstract. Japanese books are usually classified into ten genres by *Nippon Decimal Classification* (NDC) based on their subject. However, this classification is sometimes insufficient for corpus studies which describe characteristics of the texts in the book. Here, we propose a method of classifying text samples taken from Japanese books into some registers and text types. Firstly, we discuss useful criteria to describe various characteristics of the texts and propose a two-step approach for stable annotation. We then apply our method to 161 book samples from the prerelease version of the Balanced Corpus of Contemporary Written Japanese (BCCWJ), a balanced Japanese corpus comprising 100 million words developed by National Institute for Japanese Language and Linguistics. Finally, we evaluate our method in terms of stability of annotation using kappa coefficients and correlation coefficients.

Keywords: text type, register, corpus, kappa coefficient

1 Introduction

As the volume and the variations of texts compiled in language corpora are greatly increasing in recent years, providing an appropriate and informative classification code is becoming more and more important. Among many classification measures that have been proposed and employed so far, register classification, or text type annotation, has attracted a particular attention of many researchers (Biber, 1988; Tambouratzis *et al.*, 2001; Portele, 2002; Gamon 2004; Sato, Matsuyoshi and Kondoh 2008; Biber and Conrad, 2009; among others).

With appropriate classification codes, users will be able to extract on a particular set of the texts according to their objectives. For example, in many cases, usage of a word tends to depend strongly on the register and/or genre, and therefore, it is important for language researchers or learners to refer to register information of the text. In addition, appropriate register information is also crucial for creating automatic text classification systems as a source of baselines. Hence, we propose a new method of text type annotation for corpus texts taken from books.

The rest of this paper is organized as follows. Section 2 briefly reviews the style classification methods in the literature. Section 3 introduces the proposed method. Section 4 explains how our method is practically applied to a Japanese corpus called BCCWJ. Sections 5 and 6 give evaluation results and concluding remarks, respectively.

2 Style Classification

EAGLES preliminary recommendations for text typology annotation of corpora (EAGLES, 1996) makes use of three major external (E) and two major internal (I) criteria for text classification:

E.1.origin — matters concerning the origin of the text that are thought to affect its structure or content.
E.2.state — matters concerning the appearance of the text, its layout and relation to non-textual matter, at the point when it is selected for the corpus.
E.3.aims — matters concerning the reason for making the text and the intended effect it is expected to have.
I.1.topic —the subject matter, knowledge domain(s) of the text.
I.2.style — the patterns of language that are thought to correlate with external parameters.

Generally, classification with respect to subjective criteria is more difficult than objective ones. In EAGLES, E1 and E2 are objective measures, while E.3. requires subjective decisions. British National Corpus (BNC) employs attributes of expected readers and a degree of reading difficulty. In the BNC, the degree of "difficulty" was manually classified into three levels based on the annotators' quick reading, which means that this involves mostly subjective information. Information such as age and sex of expected readers, and aims of the authors, also requires subjective judgment. Furthermore, in the EAGLES example, internal criteria, I.1.topic and I.2.style, are even more difficult to find objective grounds for classification.

In this paper, we focus on style, or register classification of texts. We distinguish two issues: (1) how to choose the criteria, or parameters for the classification, and (2) how to perform the classification, along the given the criteria.

For the former issue, EAGLES (1996) reported that style is one of the most difficult criteria to define. Style is usually represented by multiple parameters in the literature, but there are no agreed standards for any one parameter. For example, EAGLES mentions that most students of language believe that a parameter of formality is required, and terms like "formal", "informal", "colloquial", etc., are used but they are not always well defined. In one of the most influential proposals (of its time), Joos (1961) sets five levels of formality, "frozen", "formal", "informal", "colloquial", "intimate". Halliday et al. (1964) suggests a primary distinction into "colloquial" and "polite", a primary distinction which is adopted in many of today's dictionaries. He also claims that the styles of discourse must be treated as a cline with categories such as "casual", "intimate" and "deferential".

For the latter issue, again, there are no standard methods for classification, except for the annotators' subjective judgment. A promising approach would be a metrical or quantitative one; for example, Biber (1988) proposed a methodology for the objective grouping of variations in English texts through statistical analyses. Establishing criteria for each parameter is necessary for corpus-based researches of texts, and to this end, extracting typical examples is a key issue offer. In the following section, we will present how these parameters should be designed, and show four criteria for the classification.

3 Method

3.1 Criteria

As a minimal first step, we propose four criteria to classify the given texts from Japanese books; (a) reading difficulty, (b) subjective-objective, (c) formal-informal, and (d) written-spoken. An important point here is that we introduce metrical measures for each criterion, so that the result of annotations can be easily used in quantitative studies and/or automatic processing. These criteria are given for each portion of the texts. A brief description of the said criteria is given below.

(a) Reading difficulty
This criterion measures the extent to which the given text is difficult for readers. We use the following metric:
1 - Texts for specialists, professional people
2 - Texts for semi-professional people
3 - Texts for generally knowledgeable people or non-specialists
4 - Texts for high-school or junior-high school students
5 - Texts for younger children

The primary factor for this metric is necessity of prior knowledge to understand the texts.

(b) Subjective or objective
This criterion measures whether the given text is written subjectively or objectively. We consider that the measure of subjective-objective is of particular importance.
1 - Very objective
2 - Rather more objective
3 - Rather more subjective
4 - Very subjective

Here, "objective" means the texts delivering the facts, observations, or logical proofs; and "subjective" means those about thoughts or impressions.

(c) Formal or informal (hard, polite, soft, or chatty)
We distinguish a degree of formal attitude (hard-soft) and politeness (polite-chatty).
1 - Very hard and polite
2 - Rather more hard and polite
3 - Rather more soft but polite
4 - Rather more soft and chatty
5 - Very soft but polite
6 - Very soft and chatty

(d) Written or spoken
In the existing corpora, book texts are categorized as written texts. However, some texts in a book like interviews and essays include colloquial expressions. Therefore, we introduce two degree of a colloquial style.
1 - Rather more a written style
2 - Rather more a spoken style

3.2 Annotation procedure

In the procedure of manual annotation, an important issue is minimizing variations among individual annotators. We have empirically found that providing typical examples is very effective for this purpose. However, finding typical examples itself is a difficult task.

Therefore, we propose a two-step approach. Firstly the annotator extracts typical examples from the corpus, and secondly annotates the classification codes shown in 3.1.

4 Experiment using BCCWJ

We applied the proposed classification method to the samples of books extracted from the Balanced Corpus of Contemporary Written Japanese (BCCWJ).

4.1 Overview of BCCWJ

BCCWJ is the first balanced corpus of contemporary written Japanese which is carefully designed from various perspectives. The corpus consists of 100 million words, and covers a wide range of text genres including books, magazines, newspapers, governmental white papers, textbooks, minutes of the National Diet, the internet text (bulletin board and blogs) and so forth. Samples taken from books, magazines and newspapers are extracted from the rigidly defined statistical populations by means of random sampling (Maekawa *et al.*, 2010). It has been designed and compiled in National Institute for Japanese Language and Linguistics since 2006, and it will be released in 2011.

4.2 Annotations

As the first step (Step 1), we first prepared 161 samples randomly chosen from the "Library Sub-corpus" of BCCWJ. Each sample consisted of a series of texts less than 10,000 characters which was randomly extracted from the pages. Then three annotators read the samples and judged the metrics mentioned in 3.1.

In the second step (Step 2), two annotators were asked to do the same, but this time, they were presented with the extracted typical examples. In this experiment, the annotation results for the same 161 reference samples were compared with those in the first step, in order to test the increase of the stability through the two-step approach.

5 Evaluation

5.1 Stability measure

We use the kappa coefficient to evaluate the stability of the register classifications. The kappa coefficient is a widely used in statistics for measuring the degree of reliability between raters (Carletta, 1996; Artstein and Poesio, 2008).

Table 1: Meaning of kappa coefficient (Landis and Koch 1977)

Kappa Statistic	Strength of Agreement
< 0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

5.2 Results and Discussions

Tables 2 through 5 list the ratios of consistent judgment, kappa coefficients, and correlation coefficients for the four criteria, respectively.

Table 2: Reading difficulty

	Step1				Step2
Annotator Pair	1–2	2–3	3–1	Average	4–5
Agreement	0.73	0.93	0.73	0.80	0.88
Kappa	0.17	0.55	0.18	0.30	0.58
Correlation	0.33	0.58	0.29	0.40	0.67

Table 3: Subjective or objective

	Step1				Step2
Annotator Pair	1−2	2−3	3−1	Average	4−5
Agreement	0.49	0.33	0.24	0.35	0.47
Kappa	0.46	0.19	0.21	0.28	0.47
Correlation	0.61	0.29	0.48	0.46	0.70

Table 4: Formal or informal (hard, polite, soft, or chatty)

	Step1				Step2
Annotator Pair	1−2	2−3	3−1	Average	4−5
Agreement	0.37	0.25	0.16	0.26	0.71
Kappa	0.24	0.21	0.07	0.17	0.67
Correlation	0.59	0.58	0.37	0.51	0.79

Table 5: Written or spoken

	Step1				Step2
Annotator Pair	1−2	2−3	3−1	Average	4−5
Agreement	0.73	0.32	0.07	0.37	0.89
Kappa	0.03	0.02	0.00	0.02	0.62
Correlation	0.13	0.06	0.02	0.07	0.63

As shown in the tables, all the consistency measures were generally low in the first step (Step 1), but they were greatly improved in the second step (Step 2). Based on this result, we expect that an appropriate set of instructions will greatly help annotators make consistent judgments. In addition, we see that the samples that yield consistent annotations in Step 1 tend to produce consistent annotations in Step 2 as well. This means that we will be able to extract effective typical examples.

Among four criteria in this experiment, the consistency measures for hard-soft, written or spoken and reading difficulty are high in that order, and yielding kappa coefficients around 0.6. As shown in Table 6 below, annotator 1 judged most samples as "rather more written" because they are actually from books, whereas annotator 3 judged many samples as "rather more spoken" because they include some story-telling parts. However, there is not a big difference between annotator 4 and 5. This shows the effectiveness of the second step.

However, for the subjective-objective measure, we only have a moderate kappa coefficient of 0.47 even in Step 2. Apparently, the subjective-objective judgment involves a non-trivial task for annotators that highly depend on personal judgment standards. We anticipate that some methodology improvement would be necessary at least for this particular criterion.

Table 6: Annotation results for the written-spoken measure

	Step1			Step2	
Annotator	1	2	3	4	5
1 − Rather more a written style	160	117	11	130	139
2 − Rather more a spoken style	1	44	150	31	22

6 Conclusion

We have proposed a set of criteria for, and a method of, register classifications for Japanese texts. The important point in the manual annotation is minimizing the fluctuations owing to individual differences. To this end, we have introduced a two-step approach, comprising a typical example extraction step, and an annotation step using the extracted typical examples as a reference. Our experimental results using the book part of the BCCWJ clearly demonstrates the effectiveness of our method; that is, in the first step without reference to the typical examples, the kappa coefficients were less than 0.3, whereas in the second step, the coefficients increased to around 0.6, except for the subjective-objective measure. We plan to extend the proposed approach to include a semi-automatic annotation step, in order to provide useful metric information about style, as part of the BCCWJ.

Acknowledgement

We thank Takehiko Maruyama and Prashant Pardeshi for their valuablel comments. This paper reports on the ongoing collaborative research project "Development of Classification Indices to Treat a Variety of Texts" at National Institute for Japanese Language and Linguistics. Part of this work is supported by MEXT Grant-in-Aid for Scientific Research on Priority Areas "Japanese Corpus".

References

Artstein, R and Poesio, M. 2008 Inter-Coder Agreement for Computational Linguistics.Computational Linguistics 34(4), 555-596

Biber, D. 1988. *Variation across speech and writing*. Cambridge: CUP.

Biber, D. and Conrad, S. 2009. *Register, Genre, and Style*. Cambridge:Cambridge University Press.

Carletta, J. 1996.Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), 249-254.

EAGLES. 1996. *EAGLES Preliminary recommendation on Text Typology.* EAGLES Document EAG–TCWG–TTYP/P, Version of Jun 1996.
(http://www.ilc.cnr.it/EAGLES/texttyp/texttyp.html)

Gamon M. 2004. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. *COLING 2004.*

Halliday, M.A.K., A, McIntosh and P, Strevens. 1964 *The linguistic sciences and language teaching.* London: Longman.

Joos, M. 1961. *The five clocks.* New York: Harcourt Brace.

Landis, J.R. and Koch, G.G. 1977. The measurement of observer agreement for categorical data. Biometrics, 33,159-174.

Maekawa, K.*et al*. 2010. Design, Compilation, and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese. *LREC 2010*, 1483-1486.

Portele, T.2002. Data-driven classification of linguistic styles in spoken dialogues. *COLING 2002*.

Sato, S, Matsuyoshi, S. and Kondoh, Y. 2008.Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus, *LREC 2008*.

Tambouratzis G., Hairetakis N., Markantonatou S. and Carayannis G. 2001. Evaluating SOM-based Models in Text Classification Tasks for the Greek Language. *Workshop on Self-organising Maps, 2001,* 11-13.