愛知教育大学教育実践総合センター紀要
第12号，pp.117～123（February，2009）

Various Problems Concerning the Construction of a WWW Japanese-Language Corpus

— The Current State and Future Prospects of Japanese-Language Corpus Research —

Hironari NOZAKI*, Kazuyuki TODA** and Kyoko UMEDA*

*Department of Information Sciences, Aichi University of Education
**Ajima Elementary School

Summary

The objective of this paper is to discuss various problems that arise when Japanese-language text that is circulating on the World-Wide Web（WWW）is utilized as a corpus. First of all, our review of previous research relating to Japanese-language corpora showed that research into the application of the WWW as a Japanese-language corpus has still not been tackled sufficiently. We then studied all of the research papers that were presented at national conventions over a two-year period for one Japanese academic society relating to information and education. As a result, it became clear that although there have been several research projects into the use of text-mining methods, there has been almost no research relating to WWW Japanese-language corpora. In the light of these findings, we considered the various problems that might arise during research into WWW Japanese-language corpora. In other words, some of the points that we need to consider include: 1）sample bias, 2）the self-images projected by authors, 3）proof of the validity of such contents, 4）the large numbers of submissions by the same people, 5）the fact that data management including the revision and update of such contents is often done at the individual level, and 6）plagiarism of written works and quoting from other sites. Thus, although sample bias does remain, we can say that the Internet gives us the first opportunity in history to accumulate vast quantities of personally-published data. We have become able to utilize the Internet both quantitatively and qualitatively as a modern intellectual resource. Once we have clarified suitable methods for using this intellectual resource as a target of research in analysis, we should be able to engage in research towards structuring a WWW Japanese-language corpus in the future.

Keywords：Japanese-Language Corpus, Loan Words, Text-mining Methods

1. Introduction

The use of unfamiliar loan words（such as “agenda” and “surveillance”）is one of the major causes of hindering the comprehension of the written word among older people and learners of Japanese. In order to assist such information-challenged people, the *Gairaigo Iikae Teian*[*1]［Proposal for Paraphrasing Loan Words］（The National Institute for Japanese Language, 2003）proposes the replacement of difficult-to-understand loan words with plain Japanese words. At the same time, together with the rapid expansion of information technology, a succession of new computer terminology（IT language）has appeared in electronic text in places such as the Internet. This IT language comprises an extremely large number of loan words that originated in English, and it often happens that such vocabulary is difficult for general readers who are not computer specialists to understand. Thus, although there are proposals for replacing difficult-to-understand loan words with plain Japanese, centered on The National Institute for Japanese Language, at the same time there is almost no research focusing on the newly-coined loan words that are proliferating rapidly on the WWW, such as IT language. In addition, even though Japanese-language text is being circulated via the Internet on a world-wide scale, not just in Japan, it cannot be said that there is sufficient research into analyzing the process of taking the actual Japanese language on the WWW as data and subjecting that data to language information processing.

Consequently, there is a strong demand for the application of Japanese-language text circulating on the WWW as a language resource, to construct a method of quantitative language processing that involves steps such as language structure analysis and meaning parsing, focusing on Japanese in electronic form（particularly text data on the Internet）. It is also thought necessary to collect and analysis examples of the usage of

vocabulary, including loan words, and research the application of such examples as a Japanese-language corpus (electronic text that can be processed by computers). Since the implementation of such research would enable clarification of the characteristics of usage and expression of vocabulary such as loan words, this research is expected to be applicable not just to linguistics, but also to Japanese-language education and information education in order to implement programs such as the selection of newly-coined loan words that appear in Japanese-language dictionaries and the extraction of collections of usage examples to assist in the comprehension of difficult-to-understand IT language concepts. This paper therefore discusses the following points, in order to implement such research: 1) an overview of previous research relating to Japanese-language corpora, 2) a study of how far research relating to WWW Japanese-language corpora has been announced at national conventions of an academic society relating to information and education, and 3) based on the above results, a clarification of various problems involved with utilizing the WWW as a Japanese-language corpus. From this we can expect to gain valuable insights into the construction of a WWW Japanese-language corpus.

2. Summary of previous research relating to Japanese-language corpora

As our information-based society expands rapidly, various different forms of Japanese-language text have been converted to electronic form and have been provided as corpora. The Aozora Bunko project is well known as a collection of literary works converted into electronic text. Literary works whose copyrights have expired have been released onto the Web, and that data is provided in plain-text format or HTML. There are also various different Japanese-language corpora that have already been developed. For example, the Sun corpus (The National Institute for Japanese Language, 2005a) was created from text of the magazine "The Sun" that was published by Hakubunkan over the period from the end of the 19th Century to the beginning of the 20th Century, during which time the written form of today's Japanese language was established. "The Sun" is interesting because it was the most widely read general-interest magazine of the time, covering a wide range of genres and featuring a diverse selection of authors. The analysis covered the editions of "The Sun" that were published from 1895 to 1928. Analysis of the Sun corpus enables studies of how Japanese moves from a literary language to a spoken language (The National Institute for Japanese Language, 2005b). In addition, Nozaki, Yokoyama, et al., (1996) have studied the usage frequencies of characters, based on a full-text database of newspaper articles. The text data that was the subject of that study was electronic text of one full year of articles from the morning and evening editions of the Asahi Newspaper, published between January 1 and December 31, 1993. It should be noted that the data also included text of 114 articles that were not within the database, which were input manually. Ultimately, approximately 110,000 newspaper articles involving more than 55,000,000 characters were subjected to this analysis, ensuring that this research had the largest number of data points within Japan at the time. Some of the results of this study were that: 1) the 1000 highest ranking kanji (Chinese-derived characters) made up about 95% of the total usage ratio, 2) the 1600 highest ranking kanji exceeded about 99% of the total, with the remaining approximately 3000 characters occupying no more than about 1% of the total, and 3) there was an extremely high correlation between newspapers of 1966 and 1993. In other words, there is a significant bias in the actual usage of kanji in newspapers, such that a very small number of frequently-used kanji are used intensively whereas a large number of low-frequency kanji are hardly used at all, and it is clear that this tendency has not changed much since approximately 30 years ago. Yokoyama, et al., (1998) also pointed out problems relating to kanji variants and JIS character codes, from a comparison of the actual newspaper pages (paper media) and the same newspaper articles on CD-ROM (electronic media). They also analyzed the appearances of "ghost characters" and characters that do not appear in dictionaries. These "ghost characters" are kanji that were created by mistake during the copying of kanji from original texts when the JIS kanji set was adopted in 1978 (Sasahara, 1997). Therefore, "ghost characters" are unfamiliar-looking characters that were mostly not even in kanji dictionaries of the time, and even their readings, meanings, and origins are unclear. Thus Yokoyama, et al., (1998) pointed out various problems that occur during the data processing of converting newspaper pages (paper media) into

electronic media. Since there is virtually no research that involves direct comparison of paper media and electronic media by referencing vast quantities of actual newspaper pages individually and in detail, to point out such problems, the research insights obtained from Yokoyama, et al., (1998) are greatly appreciated.

A further example of a corpus formed from newspaper articles is the Kyoto text corpus[*2]. Version 4.0 of the Kyoto text corpus consists of a total of approximately 40,000 sentences from all articles from January 1 to 17, 1995, and editorials from January to December, to which has been added morpheme and syntax information. This information was parsed by using a morpheme parsing system and a syntax parsing system, with the results being revised manually.

A large volume of audio language data has also been collected and converted into a form that enables its use as a corpus (Audio Resources Consortium, 2006). Such corpora are useful for promoting research and development relating to audio information processing.

Thus, although there are corpora in various different fields, such as research concentrating on Japanese-language text such as literary works or magazine and newspaper articles that were published within a specific era, or Japanese-language corpora with additional morpheme and syntax information, or audio language data turned into corpora, it still cannot be said that there is much research into the handling of Japanese-language text on the WWW as a corpus. In a similar manner to the Sun corpus, the authors of Japanese-language text circulating on the WWW are diverse and cover a wide range of genres. In addition, since this Japanese-language text is updated in real time, it is considered to be an extremely useful language resource for grasping real-life usage of modern Japanese, together with the social conditions of those times and the background of the era. It is also effective for analyzing IT language, which is full of buzzwords and newly-coined usage. In the next section, we investigate how far research relating to WWW Japanese-language corpora is being done within an academic society relating to information and education, and verify the current state of that research.

3. Current state of research into the application of Japanese-language text on the Web as a language resource.

This section presents an overview of previous research into Japanese-language text on the Web. The subject of our study was the papers presented at national conventions of the Japanese Society for Information and Systems in Education. The Japanese Society for Information and Systems in Education was founded in 1974 and approximately 1600 members have attended meetings during its 35-year history. These meetings provide opportunities for academic research and study together with the exchange of information relating to subjects such as computer usage in the educational field, and are recorded in an academic research group of the Science Council of Japan. The Japanese Society for Information and Systems in Education is therefore one of the central societies relating to Japanese education and information, after the Japan Society for Educational Technology. That is why we focused on an analysis of the Japanese Society for Information and Systems in Education.

3. 1 Subjects of study

The subjects of our study were papers presented at national conventions of the Japanese Society for Information and Systems in Education, held in 2007 and 2008. In 2007, the 32nd national convention was held over three days from September 12 (Wednesday) to 14 (Friday), at the Faculty of Engineering of Shinshu University (in Nagano City). In 2008, the 33rd national convention was held over three days from September 3 (Wednesday) to 5 (Friday) at the Faculty of Engineering of Kumamoto University (in Kumamoto City). These national conventions mainly consist of general lectures, planning sessions, and workshop. A planning session enables the research committee to determine the topic for each session and call for the submission of research papers. Other events are held too, such as panel discussions, keynote lectures, special lectures, and invitational lectures. There are also poster and demo sessions, but these two types of session were held in 2007 but not in 2008. The numbers of papers and sessions that we investigated are shown in Table 1, divided into types. The numbers of articles per session are given to two decimal places, with the third decimal place being rounded off. A comparison between 2007 and 2008 shows that the total number

of presented papers increased by about 15%. In other words: (254 − 221)/221 = 0.15. On the other hand, the number of sessions decreased. In other words: 38 − 46 = −8. The reason for the reduction in the number of sessions is considered to be mainly because there were no poster or demo sessions in 2008.

Table 1
Numbers of Papers and Sessions in 2007 and 2008

		General Lectures	Planning Sessions	Workshops	Others	Total
2007	Sessions	30	6	3	7	46
	Papers	158	37	12	14	221
	Papers / Sessions	5.27	6.17	4.00	2.00	4.80
2008	Sessions	22	6	5	5	38
	Papers	191	40	15	8	254
	Papers / Sessions	8.68	6.67	3.00	1.60	6.68

3. 2 Study procedure

(1) Extraction of text data for analysis

Our study concentrated on the presentation programs (final editions) of the national conventions of the Japanese Society for Information and Systems in Education in 2007 and 2008. We extracted the titles and subtitles of all the articles from those programs as text data.

(2) Analysis method

We analyzed the actual usage of keywords in text data created as described in (1) above, as below. The keywords used in the analysis were: "corpus," "Japanese," "text," "mining," "WWW," "Web," "Internet," "net," "network," "IT," and "ICT." These keywords are vocabulary relating to WWW Japanese-language corpora, which is the subject of these articles. Details of our methodology are as follows: (a) We extract the title and subtitle of each article that includes one or more of the keywords that are the target of this analysis, from the text created as described in (1). (b) We count the usage frequency of each keyword. (c) We extract usage examples for each keyword. (d) The authors of all the article titles and subtitles extracted in step (a) check the results visually, to confirm whether the usage frequency of each keyword is correct. (e) We also perform manual confirmation that the usage examples extracted in step (c) are valid.

3. 3 Results and considerations

The usages frequencies of the keywords that were the subject of this analysis are shown in Table 2. In addition, usage examples of keywords that had usage frequencies of 1 or more in Table 2 are analyzed in detail. These results are given in Table 3. Table 3 gives examples of the usage of the keywords and their frequencies. Each number within parentheses () in Table 3 represents the usage frequency of that example. Usage examples with no numbers in parentheses have a usage frequency of 1. Note that since the keywords "mining" and "Internet" had zero usage frequency in 2007, there are no usage examples for those words.

Table 2
Keyword Usage Frequencies of All Titles and Subtitles between 2007 and 2008

Keywords	2007	2008
コーパス	0	0
日本語	3	4
テキスト	1	1
マイニング	0	2
WWW	0	0
Ｗｅｂ	18	19
インターネット	0	2
ネット	2	2
ネットワーク	1	5
ＩＴ	1	1
ＩＣＴ	3	4

Table 3
Examples of Usage of Keywords and Their Frequencies in 2007 and 2008

Keywords	2007	2008
日本語	日本語チュートリアル，日本語ローマ字表記，日本語学習	日本語入力方式，日本語文章表現，日本語教育，日本語・中国語
テキスト	同期型遠隔協調学習用テキストチャット	テキストマイニング
マイニング	「マイニング」の使用頻度０であるため，その用例が存在しない。	学習履歴マイニング，テキストマイニング
Web	Webシステム（2），Webベース（2），Webページ（2），Webを用いた，Webアクセシビリティ，Web技術，Web教材，Web利用，Webインターフェイス，Web教育システム，Web2.0，Web学習システム，Web閲覧履歴，WebTA，Web検索	Webを利用した(4)，WebCT(2)，Webサービス(2)，マルチメディアWeb展覧会，Webポートフォリオ，WebLec8.0（システムの名称），WebOS，WebTA，Webシステム，記述式Web試験，WebELS，Webバインダ，Webリソース，Web-Based（英語論文）
インターネット	「インターネット」の使用頻度０であるため，その用例が存在しない。	インターネット市民塾，インターネット使用
ネット	ネット配信，ネットにおける	ネットいじめ，ネットにおける
ネットワーク	仮想ネットワーク	ソーシャルネットワークサービス，ワイヤレスセンサネットワーク，ネットワーク構造，大学ネットワーク，ネットワーク社会
ＩＴ	ＩＴ講習会	ＩＴリテラシー
ＩＣＴ	ＩＣＴ環境（２），ＩＣＴ活用	ＩＣＴ活用教育，ＩＣＴ活用能力，ＩＣＴ活用，ＩＣＴの役割

Note --- Usage examples with no numbers in parentheses have a usage frequency of 1.

First of all, a comparison of the data for 2007 and 2008 in Table 2 shows that the keywords for which the usage frequency increased were “Japanese,” “mining,” “Web,” “Internet,” “network,” and “ICT,” but there was no keyword for which the usage frequency dropped. One factor in the increase in the usage frequencies of these keywords is thought to be the increase in approximately 15% in the number of presented papers. The increase in the usage frequency of “mining” is worth noting. It should be mentioned that a planning session on the subject of “Data/Text Mining Intended for Educational and Study Information” was held at the national convention in 2008, at which five papers were presented. Note that since planning sessions set subjects that are expected to lead to the development of future research, it can be understood that “text mining” was a research topic that attracted attention at that conference. On the other hand, the keyword “corpus” had a usage frequency of zero in both 2007 and 2008. These results suggested that: 1) the topic of “text mining” is starting to attract attention within the Japanese Society for Information and Systems in Education, but only a few papers have been presented so far, 2) there is remarkably little research relating to Japanese-language corpora, and 3) substantially no research has been done into the application of WWW text to corpora. Consequently, there is a demand for tackling research into WWW Japanese-language corpora in the future. The next section considers various problems that might occur when applying WWW text to a Japanese-language corpus.

4. Various problems with applying WWW text to a Japanese-language corpus

From the results of the analysis of Section 3, it was clear that substantially no research has been done within the Japanese Society for Information and Systems in Education into the application of WWW text to Japanese-language corpora. In this section, we would like to verify various problems relating to turning Japanese-language text on the WWW into a corpus, and obtain some insights to contribute to the development of research in the future.

As is well known, there is a growing flood of

information on the Internet. The advances in multimedia functions of personal computers have led to the presence of huge amounts of different types of data on the net; not just Japanese-language text, but also in other forms such as video and audio data. There are also many applications that ought to be able to make use of such information. However, there has been insufficient investigation of the methodology of how to incorporate such information. Consequently, this paper discusses various problems that might occur when tackling research into the collection and analysis of the flood of Japanese-language text that is on the Internet, to create a useful WWW Japanese-language corpus. As a precondition to promote this debate, we focused on research into Japanese-language text alone, from amongst the huge variety of information that is available on the WWW. In other words, multimedia contents such as video or audio data were outside the scope of this analysis. Audio data in particular can be applied to a Japanese audio corpus, but information technology that can handle large quantities of audio data efficiently is not yet up to the task. It is therefore considered valid to concentrate on just Japanese-language text in the current state of the art.

It is first necessary to review how we can handle this flood of data. As a general rule, we cannot determine the background of Japanese-language text on the Internet, such as the gender or personality of each author and the reliability and credibility of contents. Thus, when analyzing Japanese-language text data on the WWW, there are various factors that skew the data such as: 1) sample bias, 2) the self-images projected by authors, 3) proof of the validity of such contents, 4) the large numbers of submissions by the same people, 5) the fact that data management including the revision and update of such contents is often done at the individual level, and 6) plagiarism of written works and quoting from other sites. As a general rule on the web, the audience often cannot identify the authors of text. This apparent anonymity is thought to be one major factor in a deterioration of morals of authors. Even when there is no malicious intent, it is considered difficult to rigorously determine the correctness or validity of statements based on misunderstandings or errors of the authors, or statements written in hobby or fun categories. Since the Internet features such a huge diversity of authors, the opinions of all levels of the public are reflected widely on it, but at the same time the contents of their statements are a mixture of good and bad.

In order to focus research onto such Japanese-language text, we must consider the following points: First of all, 1) since it is impossible to identify individual authors, this data is unsuitable for investigating distributions of opinions and the spread of debate. 2) Since this is a biased sample, a very small number of people account for a large proportion of the submissions. In other words, an extremely large number of people are just seen and do not deliver. Although such a sample bias does remain, 3) the Internet has enabled people from a wide range of levels to post information. 4) Large amounts of intellectual resources are often available on the Internet. 5) The Internet gives us the first opportunity in history to accumulate vast quantities of personally-published data.

This means that it has become possible for us to use the Internet both quantitatively and qualitatively as an up-to-date intellectual resource. However, various reservations are necessary for using such data as the subject of study in analyses. As described above, although the samples are biased, they are an effective means of accessing variations in the individual thoughts of a diverse range of people. To determine whether it is possible to handle research into Japanese-language text data that is on the WWW as a corpus, it will be necessary to conduct further review of specific research methods in the future.

As computers have developed, it has become possible to save huge quantities of information. However, methods of handling such information in a unified manner remain insufficient, and there is a strong demand for the new construction of such methods. In conclusion, we list some challenges that should be tackled in the future: 1) Consider what we can say from the flood of Japanese-language text on the Internet. 2) Support comprehending and writing Japanese-language text, using a corpus. 3) Video and audio data on the WWW was outside the scope of analysis of this study, but such data should also be the target of study in the future.

5. Conclusions

In this paper, we discussed the necessity of constructing a WWW Japanese-language corpus. More specifically, during an overview of previous research relating to Japanese-language corpora,

we discovered that there has been insufficient research into the formation of a corpus from the large quantity of Japanese-language text that exists on the WWW. We then investigated all of the papers that were presented during two full years of national conventions of the Japanese Society for Information and Systems in Education. As a result, it was demonstrated that although there has been virtually no research relating to corpora, research using text mining methods are being pursued. Finally, we discussed various problems that could be encountered during the application of WWW text to a Japanese-language corpus, in the light of these findings. The construction of a WWW Japanese-language corpus can be expected to involve not just Japanese language research, but also assistance with the comprehension of difficult-to-understand IT language concepts, assistance with the comprehension of Japanese writings, and clarification of the roots of newly-coined vocabulary such as buzzwords. In the future, it will be necessary to construct specific methods for converting WWW text into Japanese-language corpora.

ACKNOWLEDGMENT

We would like to express our gratitude for the partial support for this research from the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research（B）, 19300280.

REFERENCE

[1] National Institute of Informatics（2006）Speech Resources Consortium, http://research.nii.ac.jp/src/

[2] National Institute for Japanese Language（2005a）*Taiyo kopasu* [the Sun corpus], Hakubunkan, Tokyo

[3] National Institute for Japanese Language（2005b）*Zasshi Taiyo niyoru kakurituki gendaigo no kenkyu* [A study of early modern Japanese language on the Sun corpus], Hakubunkan, Tokyo

[4] Nozaki, H., Yokoyama, S., Isomoto, Y., and Yoneda, J.（1996）*Moji shiyo ni kansuru keiryoteki kenkyu* [A study of character frequency --From the view point of Japanese language education], Japan Journal of Educational Technology, **20**(3), 141-149

[5] Yokoyama, S., Sasahara, H., Nozaki, H., and Long, E.（1998）*Shinbun denshi media no kanji* [A study of kanji in electronic newspaper media], Sanseido, Tokyo.

[6] Aozora bunko（2007）http://www.aozora.gr.jp

[7] Sasahara, H.（1997）*Jitai ni syoujiru guuzen no iti*, Japanese Linguistics 1, National Institute for Japanese Language, 7-24, Tokyo

[8] National Institute for Japanese Language（2003）*Gairaigo Iikae Teian* [Proposal for Paraphrasing Loan Words], Tokyo

NOTES

1. The National Institute for Japanese Language provides tables of loan words that may not be understood easily and correctly, with suggested paraphrases in Japanese words. The lists of loan words provide paraphrases as well as guidelines for their usage. For some loan words, multiple sample paraphrases are provided, to accurately reflect the contexts of their use. These loan words lists are called "*Gairaigo Iikae Teian*" in Japanese. http://www.kokken.go.jp/gairaigo/（link to Japanese page）
2. Kyoto text corpus Version 4.0 http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/corpus.html