Kotonoha, the Corpus Development Project of the National Institute for Japanese Language

Kikuo Maekawa

Department of Language Research
National Institute for Japanese Language
kikuo@kokken.go.jp

Abstract

The aim of the Kotonoha project consists in providing a series of corpora that covers the full range of present-day Japanese language both spoken and written. So far, NIJL has released two component corpora of Kotonoha; the Corpus of Spontaneous Japanese (CSJ) and the Taiyo Corpus. CSJ is a large-scale richly annotated corpus of spontaneous monologue developed originally for automatic speech recognition of spontaneous speech, but is also designed for the study of language variation. On the other hand, Taiyo Corpus is a full-text corpus of a general interest magazine Taiyo, which was very popular among the educated people in prewar times. The corpus consisted of the XML-formatted text of 60 volumes covering the period of 1895-1925. Results of analysis using CSJ and/or Taiyo will be presented briefly. As for the coming years, the first corpus to be compiled in Kotonoha project is a balanced corpus of the present-day written Japanese. Our current plan is to provide annotated text of 50 million words in the coming five years. It might also be possible to double the amount of the corpus if we are successful in obtaining a national grant that we submitted recently.

1. Introduction

As shown by the papers of this symposium, study of the Japanese language lags behind as long as modern corpus linguistics is concerned. It is widely acknowledged by those who work in the field that one of the fundamental problems in Japanese corpus linguistics is the lack of a so-called 'balanced' corpus, a corpus that represents the whole range of the target language in statistically unbiased manner.

To fill this lag, National Institute for Japanese Language (NIJL) is going to launch a new corpus creation project this spring. Adopting an archaic Japanese meaning 'word' and/or 'language' the project is named *Kotonoha* project.

The first part of this paper is devoted to the overview of *Kotonoha* project. As will be shown, *Kotonoha* is not a single corpus but a cover-term for a set of corpora that, in their entirety, cover the whole range of modern Japanese. The second part of the paper is devoted to the presentation of the design and analysis of two existing component corpora of *Kotonoha*: CSJ and *Taiyo*. The last part of the paper deals with design issues of a balanced corpus of the present-day written Japanese; the first component corpus of *Kotonoha* to be compiled.

2. *Kotonoha*

As stated earlier, *Kotonoha* is a cover term for a series of corpora that covers the whole range of modern Japanese. The range covers both spoken and written Japanese beginning from the year of Meiji Restoration (*i.e.*, after 1868).

Figure 1 shows important component corpora of *Kotonoha*, existing as well as to be created. The abscissa of the figure represents schematically time after 1868. The ordinate makes distinction between the written and spoken varieties as well as many sub-varieties in each category. Arrows in the bottom of the figure stand for three epochs of great change in the history of modern Japanese, namely, the shift from classical literary writing style (*Bungotai*) to modern colloquial writing style (*Kougotai*), proclamation of the present-day Kana orthography (*Gendaikanazukai*) in 1946, and, beginning of computer processing of Japanese characters as exemplified by the announcement of the first JIS Kanji code in 1978.

Ellipses in the figure stand for existing corpora, i.e., *Taiyo* and *CSJ*. These will be discussed in more details in the following sections. Squares in the figure are the corpora to be created. Located on the right upper corner is a balanced corpus of the contemporary written Japanese (BCCWJ, hereafter). This is the corpus that NIJL will develop in the coming five years, which will be the theme of Makoto Yamazaki's talk.

Two squares with dotted outlines are corpora that might be created after the development of BCCWJ (after 2011 according to the current time table). One of them, located under the CSJ, is a corpus of natural conversational discourse. This corpus will fill the gap left by the CSJ in the genre of present-day spoken Japanese.

The other dotted square is a corpus covering the period of time between the *Taiyo* and BCCWJ. This corpus is important because this period include the discontinuity of the Japanese society (and language) caused by Japan's loss of World War II.

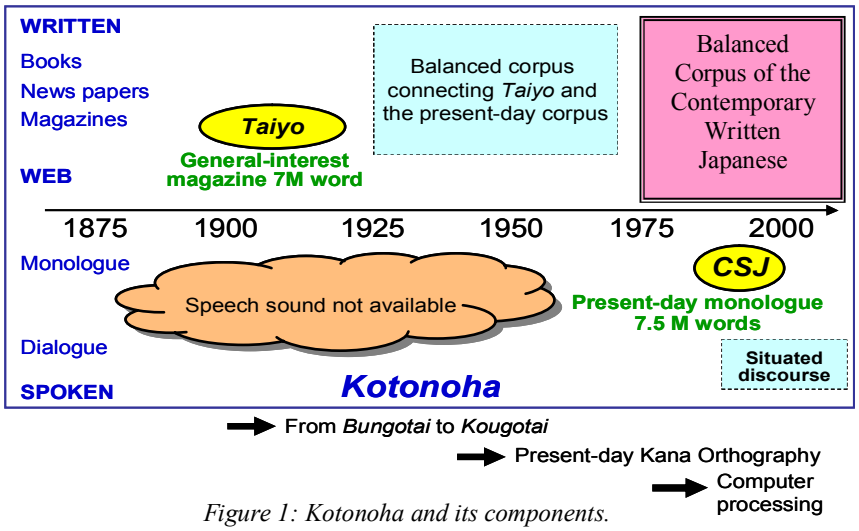

Figure 1: Kotonoha and its components.

One crucial difference between this corpus and BCCWJ is the inclusion of WEB text in the latter. BCCWJ can be regarded as a corpus of written Japanese in the age of computer and internet.

Lastly, there is a considerable period of time where creation of a full-fledged spoken corpora is virtually impossible: the period prior to the diffusion of magnetic tape recorder in the 1960s.

3. Existing corpora

3.1. Corpus of Spontaneous Japanese

CSJ, or *Corpus of Spontaneous Japanese*, is a richly annotated large-scale corpus of spontaneous monologue designed primarily for the study of automatic speech recognition and the study of language variation [1-3].

More than 660 hours of the speech signal contained in the corpus have been transcribed and POS annotated. Detailed annotation has been conducted on a subset of the corpus called CSJ-Core. See table 1.

Table1: Annotation of CSJ

ANNOTATION	Whole CSJ	CSJ-Core
Speech signal	✓	✓
Speaker info	✓	✓
Transcription	✓	✓
POS analysis	✓	✓
Clause boundary Info	✓	✓
Impressionistic rating	✓	✓
Segmental label	N.A.	✓
Intonation label	N.A.	✓
Dependency analysis	N.A.	✓
Topic boundary info*	N.A.	✓

* Available only for a subset of the CSJ-Core.
N.A. means 'Not Available.'

'Impressionistic rating' in table 1 requires special mention here. Because talks in the CSJ differ considerably with respect to the way they were spoken, they were evaluated by human raters at the time of recording with respect to various subjective dimensions, including speaking rate, speaking style, spontaneity of talks, use of prepared text, etc. Speaking style, for example, is evaluated using the scale of 1 to 5, higher the number more formal the speech. The usefulness of impressionistic rating score will be demonstrated later in this section.

Use of CSJ in spontaneous speech recognition increased recognition rate drastically. Word recognition rate jumped up from 45% to 80% as the combined effect of new language- and acoustic- models learnt from CSJ and improvements in pattern recognition techniques like speaker adaptation [4].

3.1.1. Word coalescence

CSJ has also proved to be effective for the study of language variation. Figure 2, taken from an unpublished manuscript of the present author, shows the result of so-called 'decision tree' analysis of word coalescence. The coalescence analyzed here is that of /de/ and /wa/ into /zya/. The most important factor, as represented by the root node of the tree, is the POS of /de/ (either case particle or adverbial form of auxiliary /da/). While the rate of coalescence is low （about 2%） in the case of particle, it is high （about 43%） in the case of auxiliary.

Social factors come into the picture at the second highest node level. It is the so-called impressionistic rating score of speaking style that constitutes the node immediately below the node of particle, and, it is the type of monologue talks (either Academic Presentation Speech or Simulated Public Speaking) that constitutes the node below the auxiliary node. Lower score of

speaking style (i.e. less formal style) and public speaking result in higher rate of coalescence. This kind of quantitative analysis on the factor of linguistic variation can be performed for many cases, because CSJ is richly annotated in view of its use for variation studies.

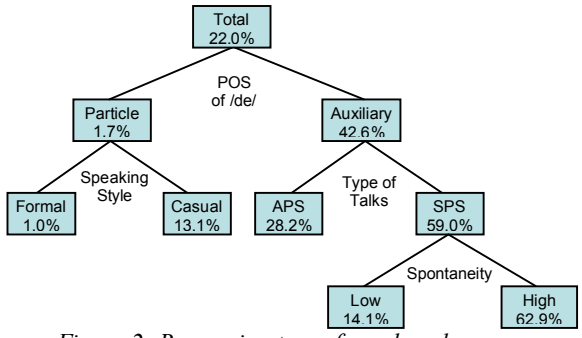

Figure 2: Regression tree of word coalescence /de/+/wa/ => /zya/. Digits show the rate of coalescence.

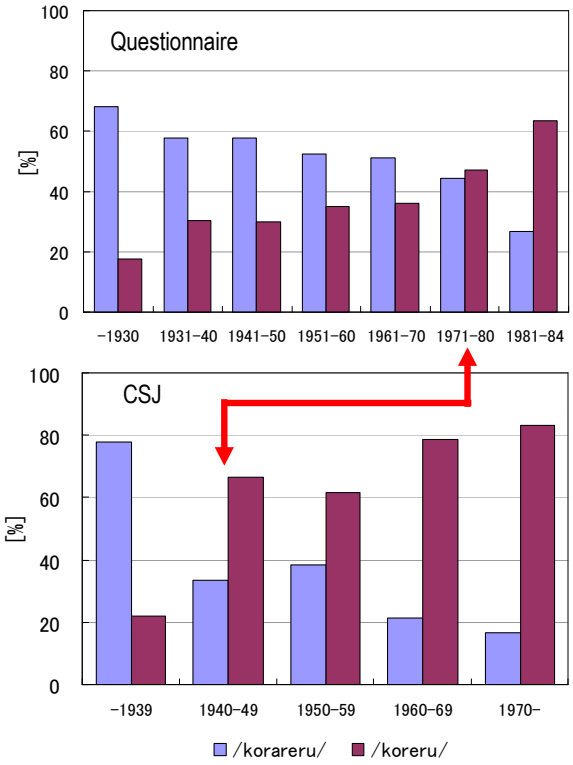

Figure 3: Comparison of questionnaire survey and analysis of CSJ. Potential form of a verb /kuru/.

3.1.2. *Comparison with questionnaire survey*

Figure 3 compares the result of questionnaire survey done by the Japanese government's Agency of Cultural Affairs in 2001 and the analysis of CSJ data collected mostly in the years 1999-2002, with respect to the ongoing morphological change in the potential form of verb /kuru/ (*'come'*) [5].

As is well known, innovative potential form of /koreru/ has been becoming more and more popular during the last two centuries or so. As the result of the historical change, innovative /koreru/ is much more popular than traditional /koreru/ especially among the young speakers.

Both panels in Figure 3 show the usage rate of the innovative form (ordinate) as a function of the speakers' birth year (abscissa). Here, I want to attract readers' attention to the timing of reversal, i.e., the point on abscissa when innovative form became more frequent than the traditional form for the first time (indicated by the arrow in the figure).

The questionnaire data indicates that it is the group of subjects born in the 1970s that constitutes the timing of reversal. The CSJ data indicates, on the other hand, that the timing of reversal is found as early as in the group of speakers born in the 1940s. Why this difference?

Needless to say, it is the CSJ data that reveals the real speech behavior of subjects. The CSJ data is the direct observation of speech, while the questionnaire data is the subjects' interpretation of their own speech behavior, which can be biased easily by things like consciousness of linguistic norm.

3.1.3. *Boundary pitch movements*

The next example is on intonation. As is well known among phoneticians, intonation of Japanese utterances can be described in terms of the characteristics of basic unit of prosody known as the accentual phrase. The final part of an accentual phrase (its right-hand edge) is especially important for the description of Japanese intonation, because the right edges are often marked by boundary pitch movements, or BPM [6], that play crucial role in the transmission of pragmatic and/or paralinguistic meaning [7].

Table 2 shows the frequencies and rates of BPM recorded in the CSJ-Core. Rows and columns of the table correspond respectively to the type of talks and type of BPM. As for the type of BPM, three main types were analyzed here, i.e. H% (rising tune), HL% (rising-falling tune), and, LH% (rising tune with sustained low pitch at the beginning). The last column (L%) shows the cases where no BPM occurred.

Table 2: Frequency of three BPM in the CSJ-Core

Talk type	H%	HL%	LH%	L%
APS	19,647 (25.4%)	3,041 (3.9%)	125 (0.2%)	54,605 (70.5%)
SPS	11,148 (13.8%)	6,997 (8.7%)	259 (0.3%)	62,127 (77.1%)

Table 2 suggests the possibility that monologue types, APS and SPS, could be discerned, at least to some extent, by their intonation characteristic, because there is near trading relationship between the rates of H% and HL%.

The validity of this hypothesis can be confirmed visually by making the scatter plot of all APS and SPS files with respect to their rates of H% and HL%.(See figure 4).

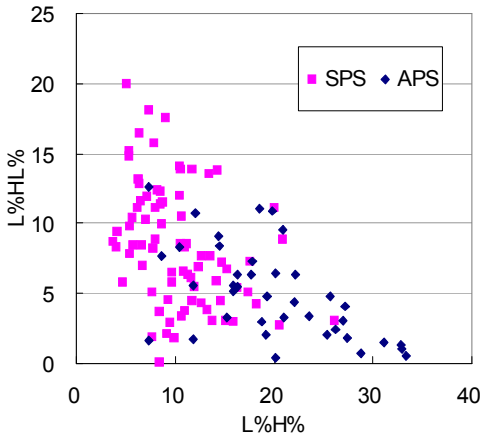

Figure 4: Scatter plot of APS and SPS talks in the CSJ-Core with respect to the rates of H% and HL%.

The last example is about the variation of particular BPM. Fine phonetic observation of the HL% BPM tells us that there is variation with respect to the temporal alignment between the pitch peaks and the texts.

Suppose, for example, an HL% BPM mark the end of an accentual phrase /anode'Htaga/ (where /ano/ means 'that', /deHta/ means 'data' and /ga/ is a case particle of AGENT. Note in passing the apostrophe in /de'Hta/ stands for a lexically specified pitch accent). In most cases, the peak is aligned to the last syllable of the phrase, i.e. /ga/, resulting in a rising-falling local intonation occurring within a single syllable of /ga/.

There are, however, cases where the peak is aligned to the penult syllable, i.e. /ta/. The latter case is labeled as an instance of 'PNLP' (*Penult Non-Lexical Prominence*) variant of the HL% BPM. See figure 5 for comparison of the two variants.

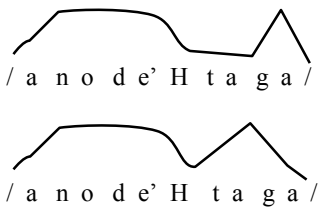

Figure 5: Schematic comparison of the pitch shapes of ordinary (top) and PNLP (bottom) variants of HL%.

Table 3 shows the frequencies of ordinary and PNLP variants in APS and SPS. It is noticeable that the rate of PNLP is higher in APS than in SPS.

In addition, figure 6 shows the relationship between the rate of PNLP variant and the impressionistic rating score of the preparedness of talks. Preparedness 0 means that speakers do not rely upon any prepared manuscript in their talks, and, preparedness 2 means that speakers are reading prepared manuscript. Preparedness 1 is intermediate. This figure suggests the interpretation that the PNLP variant is used primarily in the speech of reading style. Detailed examination of this hypothesis is an interesting research topic of the pragmatics of Japanese prosody.

Table 3: Frequency of ordinary and PNLP variants.

Talk type	Ordinary	PNLP
APS	2,312 (76.0%)	729 (24.0%)
SPS	6,605 (94.4%)	392 (5.6%)

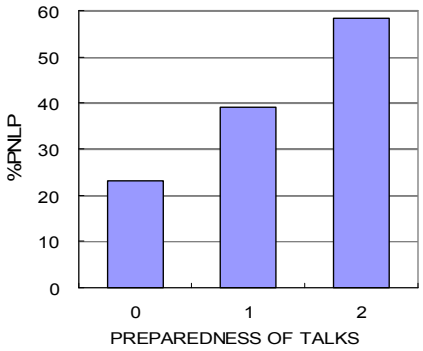

Figure 6: Correlation of the ratio of PNLP variant and impressionistic rating score of the preparedness of talks.

Since its public release in June 2004, NIJL shipped more than 270 sets of CSJ. Users include universities, national laboratories, laboratories of private companies, and individuals who are interested in speech. These users transverse both engineering and humanities domains. To our knowledge, more than 440 papers have made reference to or analyzed CSJ.

3.2. Taiyo Corpus

The second corpus currently available from NIJL is the *Taiyo Corpus* [8, 9]. This corpus was constructed for the study of modern Japanese in the period when modern colloquial writing style (*Kougobun*) was established.

Taiyo Corpus is a text corpus of written Japanese as represented by the articles of a magazine *Taiyo* published by *Hakubunkan* publishing. The magazine

was selected as the source of the corpus for two reasons. Firstly, *Taiyo* was a magazine of general interest. Its articles covered very wide range of topics. Secondly, *Taiyo* was one of the best selling magazines of the time, read by wide range of readers.

The *Taiyo Corpus* consists of 3400 articles written by about 1000 writers published in the years 1895, 1901, 1909, 1917, and, 1925. The total amount of the corpus is 14.5 million characters, corresponding to roughly about 7 million words.

Figure 7 shows the number of articles as classified by the NDC (Nippon Decimal Classification) system, which is the standard classification system of the contents of library. This figure shows that *Taiyo Corpus* has wide coverage of texts encompassing all primary categories of NDC.

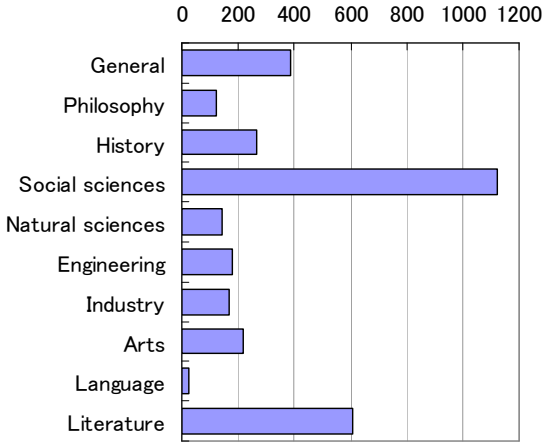

Figure 7: Classification of articles by means of NDC.

Like CSJ, the whole texts are formatted by means of XML with respect to the information like authors, writing styles, errata in the original texts, authors of the quoted text, and so on. POS analysis, however, was not provided. POS analysis of the *Taiyo* texts is extremely difficult due mainly to the complexity of writing styles. Both literary and colloquial writing styles coexisted throughout the years 1895-1925 as shown in figure 8 [10].

This figure is based upon the binary classification of the writing style of articles, and the classification is done based upon the stylistic characteristics of auxiliary verbs in the predicate phrases. This is an extremely simplified classification.. In reality literary and colloquial expressions coexisted even in a single article, thereby making automatic POS analysis almost impossible.

Lack of POS information does not deteriorate the value of the corpus considerably. Here are some very interesting results obtained by the analysis of *Taiyo Corpus*.

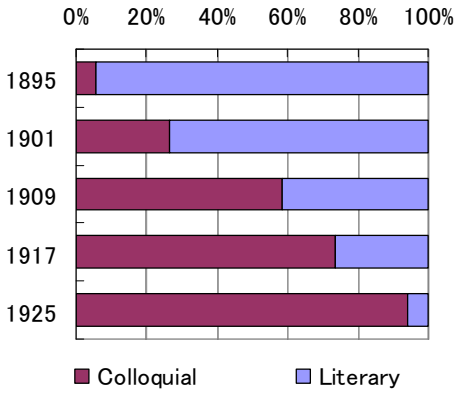

Figure8: Ratio of articles written in colloquial and literary writing styles.

3.2.1. Dakuten

Figure 9 shows the change in the usage rate of '*dakuten*' as a function of time and writing style [11]. *Dakuten* is a diacritic used with Japanese *Kana* letters to denote that the *Kana* involves voiced consonant. For example, kana letters く and さ stand respectively for /ku/ and /sa/, while ぐ and ざ with *dakuten*, stand for /gu/ and /za/ respectively.

Figure 9 shows that the use of *dakuten* was not consistent in the end of the 19th century, especially in articles written in literary style, but the inconsistency disappeared quickly as time went by.

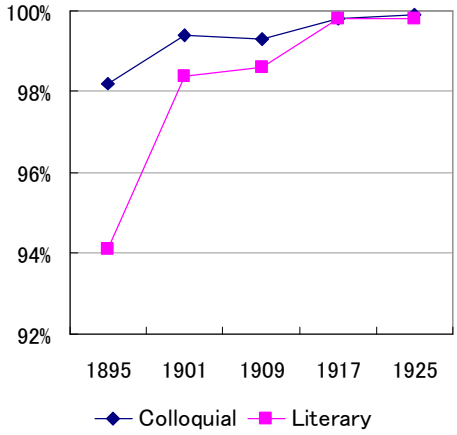

Figure 9: Usage rate of 'dakuten.'

3.2.2. Potential forms of Sino-Japanese verbs

We have already seen in 3.1.2 that there is variation in the potential form of verbs. Figure 3 was concerned with conjugational morphology of verbs, but this is not the only source of variation. At the time of *Taiyo*, considerable variation was observed in verbs of Sino-Japanese verb stem with respect to the choice of verb-

making suffixes as shown in table 4, where 'X' stands for Sino-Japanese verb stem. Potential forms are classified into three types according to the last (head) element of the suffixes in this table.

TaAble 4: Potential forms of Sino-Japanese verbs

Type	Potential Form
ATAU	X suru koto atau
	X suru atau
	X shi atau
ERU	X suru koto o eru
	X suru o eru
	X shi eru
DEKIRU	X suru koto ga dekiru
	X ga dekiru
	X dekiru

Figure 10 shows the frequency of DEKIRU, ATAU, SHIERU, and, SURUOERU; the last two being the sub-types of ERU listed in table 4. This figure also shows the percentage of the articles written in colloquial writing style as the right-hand ordinate.

Frequency of DEKIRU, which can be regarded to be the standard form of the present-day Japanese, increased monotonically throughout the years 1895-1925, presumably as the result of increase in the rate of colloquial articles [12].

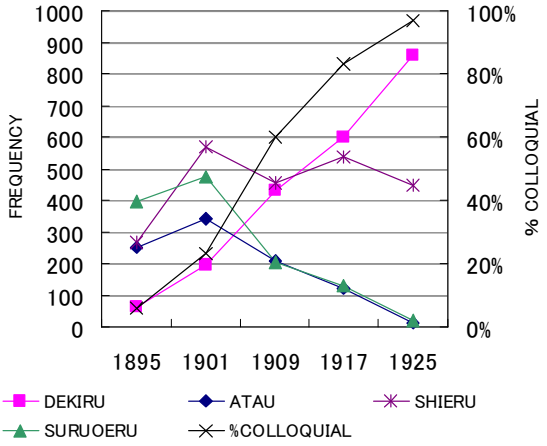

Figure 10: Frequency of various potential forms of Sino-Japanese verbs.

3.2.3. Survival race of words

In Meiji era, number of Sino-Japanese words exploded due mainly to the modernization and westernization of Japanese society. At the period of time covered by *Taiyo Corpus*, however, new words were in the process of shakeout.

Figure 11 shows 6 words, 4 Sino-Japanese and 2 Japanese native, that stand for the concept of being 'excellent.' Traditional Japanese word '*sugureru*' was always the most frequent, but the word of the second highest frequency changed frequently in the years 1895-1909, ending in the survival of '*yuushuu.*'

Based upon the analysis of collocation patterns, Tanaka found subtle but consistent semantic difference between the usages of '*sugureru*' and '*yuushuu*'[13].

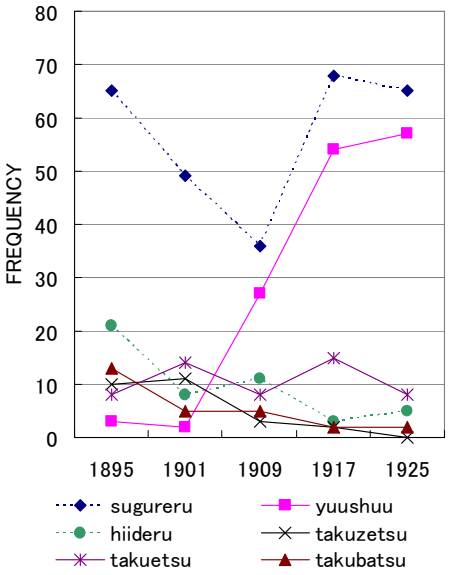

Figure 11: Survival of words for "excellent".

4. Design issues of the BCCWJ

In 2005, we constructed a small balanced corpus of about one million words, as a pilot evaluation corpus of the BCCWJ. Some of the issues that we discussed during the course of its construction are presented below.

4.1. Necessity of sampling

There are people who think that sampling of text is an obsolete technique, because nowadays we can make access to very large amount of electronic text in the Internet.

It is true that the Internet has changed drastically the way we collect linguistic data. It may also change the way we construct corpora. But data sampling is still an indispensable part of corpus design as long as we plan to make a balanced corpus. Statistical analysis of linguistic data collected by simple web search may run the risk of heavy bias by the factors like frequency of topic and the ordering of URLs by the search engine (of which detailed algorithm is not usually available).

The Web as Corpus project reported by Marco Baroni in this symposium [14] is worth special mention; this is an attempt to construct a sort of balanced corpus out of the web data. This is certainly

one of the most probable futuristic views of language corpus in the decades to come.

Nonetheless, I would like to point out that a balanced corpus constructed by traditional method is indispensable in Japanese, because traditional corpus is necessary for the evaluation of web-based balanced corpus anyway, and Japanese is lacking traditional corpus.

4.2. Corpus size

The second question is whether hundred million words are good enough as the size of BCWL. The answer to this question depends clearly upon the usage of corpus, but there are undoubtedly cases where hundred millions is not sufficient.

Dictionary compilation should be one such case. Based upon the results of various word surveys done by NIJL (see next section), we suppose that the number of different words (types) obtained by hundred million word corpus should be about in the order of hundred thousands. One hundred thousand approximate the size of compact-size Japanese dictionaries like *Sanseido Kokugo Jiten* (5th ed., 76,000 entries) and *Shueisha Kokugo Jiten* (1st ed., 92,000 entries).

But the fact is we need much larger corpus to compile dictionaries of this size, because most words (usually more than the half of all types) occur just once in corpus. History of scaling-up of English corpora is reviewed in the paper of Stephen Bullon in this symposium [15].

In purely academic research, we would also need very large corpus if we conduct a study like [16] where occurrence probability of so-called 'deviant' and 'agrammatical' sentences like "Colorless green idea sleeps furiously" and "Furiously sleep ideas green colorless" were estimated based upon bigram model with hidden variables.

The size of BCCWJ was set for no scientific reason. It was rather an estimation of the upper bound of our budget (and, to a lesser extent, the size of BNC) that had overriding influence on our thinking.

I am completely aware that some users of the corpus will start complaining about the corpus size as soon as the corpus becomes available.

There are two practical solutions to the problem. Firstly, we will continue enlarging the corpus even after we achieved the goal of one hundred million, with probably the rate of one million words per year. Secondly, we may start compiling web-based balanced corpus in the future.

4.3. Sampling and sample length

The issue of sampling is worth serious discussion, because the notion is central to the conception of balanced corpus. As will be shown by Yamazaki's paper in this symposium, the sampling method of BCWL is statistically rigid in the sense that it is basically a random sampling from a population (and it is especially true with the 'fixed-length' sample part of the corpus as opposed to the 'variable-length' part). This reflects the tradition of Japanese quantitative lexicology developed by NIJL.

From a point of view of the sampling theory of statistics, the best representativeness of a corpus is achieved when each sample in a population has the same probability of being drawn. This is the well-known principle of random sampling.

As long as I know, however, no language corpus has ever been compiled based strictly upon the principle of random sampling. In the case of the famous *Brown Corpus*, the corpus consists of 15 genres like 'Press: reportage', 'Press: editorial', 'Religion', 'Skills and hobbies' and so forth. According to the manual of *Brown Corpus* [17], the selection of text in each genre was nearly random, but the setting of the 15 genres and the allocation of the number of samples to each genre were human choices based upon intuitive evaluation of the linguistic importance of the genres. Text selection procedure of the *British National Corpus* seems to be much less random than that of the *Brown Corpus*.

On the contrary, NIJL researchers have been using almost exclusively random sampling as a basic technique of their statistical "word surveys" during the last 50 years. The first NIJL word survey based upon random sampling was conducted in 1952 using the population of '*Syuhu no tomo*' magazines published in 1950. About 150,000 running words were extracted from the population of 900,000. The survey resulted in a word frequency list consisting of about 27,000 different words (types).

In 1957, a famous survey known as the survey of "90 magazines" was conducted, in which 530,000 samples were extracted from the population of 160 million words, resulting in the frequency list of about 40,000 different words.

In the years that followed, similar surveys were conducted repeatedly for various populations: newspapers, women's magazines, junior high school textbooks, high school textbooks, TV programs, and, 70 magazines.

At this point, astute readers might want to know if the samples drawn in these surveys could be used as samples for *Kotonoha*. Unfortunately, it is impossible, because sample length of the past NIJL surveys were too short to be included in *Kotonoha* or whatever language corpus. In the case of the latest survey (known as the survey of "70 magazines" published in 1994), the averaged sample length was 70.3 character long, which corresponds to about 39 morphemes on average.

The past NIJL surveys used short samples in order to get as much information as possible of lexical diversity. Figure 11 shows the result of simulation study by Masaya Yamaguchi of NIJL, showing the relationship between the sample length (expressed in

terms of the number of morphemes in a sample) and the number of different morphemes (types) included in the whole samples. The size of population was 9,850,827 morphemes for books (153 titles), and 5,480,951 for magazines (50 titles).

Sample length 1 is the extreme case of short sample length where each sample consists of a single morpheme, and sample length 919,700 is the other extreme, where only a single sample of very long size was drawn from the population. There is clear negative correlation between the sample length and number of types. Also, it can be seen from the figure that number of types diminishes quickly at the beginning of abscissa, but the change is not abrupt when the sample length is longer than 500.

Needless to say, very short sample ran the risk of opaque context. So it is important to find a point of equilibrium between the requirements of sampling and those of corpus usage. This is something for which we spent a lot of time, but I want to leave the discussion of this issue to Makoto Yamazaki.

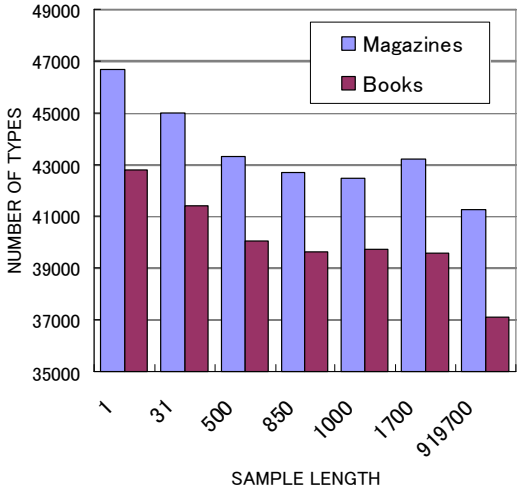

Figure 11: Relationship between sample length (N morphemes) and the number of type words.

5. Concluding Remark

This April, we will start the construction of BCCWJ which will be the third component corpus of *Kotonoha*. But the design of the BCCWJ has not been completely fixed. We would very much like to listen to any constructive comments regarding the design of BCCWJ and *Kotonoha*. Please contact us with the email address on the title page of our papers.

ACKNOWLEDEGEMENT

I'm very grateful for my colleagues Makoto Yamazaki, Makiro Tanaka, Masaya Yamaguchi, and Caroline Menezes for their help in preparing this manuscript.

6. References

[1] Maekawa, K. (2004). "Nihongo hanashikotoba koopasu no gaiyou." In *Nihongokagaku*, 15, pp. 111-133.

[2] Maekawa, K., H. Kikuchi and W. Tsukahara (2004). "Corpus of Spontaneous Japanese: Design, Annotation and XML Representation", *Proceedings of the International Symposium on Large-scale Knowledge Resources*, Tokyo Institute of Technology, pp.19-24.

[3] http://www2.kokken.go.jp/%7Ecsj/public/index.html

[4] Furui, S. M. Nakamura, T. Ichiba, and K. Iwano (2005). "Analysis and recognition of spontaneous speech using Corpus of Spontaneous Japanese." *Speech Communication*, 47 (1/2), pp. 208-219.

[5] Maekawa, K.(2005)."Quantitative analysis of word-form variation using a spontaneous speech corpus", *Proceedings of Corpus Linguistics*, Birmingham.

[6] Maekawa, K., H. Koiso, H. Kikuchi and K. Yoneyama (2004). "Use of a large-scale spontaneous speech corpus in the study of linguistic variation", *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, pp.643-646.

[7] Maekawa, K (2004). "Production and Perception of 'Paralinguistic' Information", *Proceedings of International Conference: Speech Prosody 2004*, Nara, pp.367-374.

[8] Nt'l. Inst. Jap. Lang. (2005). *Zasshi taiyo niyoru kakuritsuki gendaigo no kenkyuu*. Hakubunkan-shinsha, Tokyo (NIJL Research Report No. 122).

[9] Tanaka, M. (2005). "Gengosiryoo toshiteno zasshi taiyo no koosatsu to taiyo koopasu no sekkei." In NIJL (2005).

[10] http://www.kokken.go.jp/lrc/

[11] Kondo, A.(2005). "Dakuten shiyooritsu kara miru dakuon hyouki." In NIJL (2005).

[12] Ogiso, T. (2005). "Kango sahen doushi no kanou no katachi." In NIJL (2005).

[13] Tanaka, M. (2005). "Kango 'yuushuu' no teichaku to goi keisei." In NIJL (2005).

[14] Baroni, M. (2006). "Building general- and special-purpose corpora by Web crawling." *This proceedings*.

[15] Bullon, S. (2006). "The use of corpora in pedagogical lexicography." *This proceedings*.

[16] Pereira, F. (2000). "Formal grammar and information theory: Together again?" *Philosophical Transactions of the Royal Society*, 358(1769): pp.1239-1253.

[17] http://khnt.hit.uib.no/icame/manuals/brown/

[18] Yamazaki, M. "Design of NIJL Balanced Corpus of Contemporary Written Japanese." *This proceedings.*