Analysis of Language Variation Using a Large-Scale Corpus of Spontaneous Speech

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

ABSTRACT

Large-scale corpus of spontaneous speech can be a powerful tool for the study of language variation. Moreover, given that the corpus is publicly available, corpus-based analysis could open up the possibility of follow-up analysis in this area of linguistic study. Generally speaking, follow-up study is highly desirable in sciences but so far it has been virtually impossible in the area of socio-linguistics due to the lack of shared corpus. In my talk, I will present some results of the analyses of the *Corpus of Spontaneous Japanese* (CSJ) that we developed in the years 1999-2003.

CSJ is a large, richly annotated corpus of spontaneous speech of the present-day Japanese (http://www2.kokken.go.jp/~csj/public/index.html), containing more than 660 hours of speech uttered by more than 1400 speakers. This corpus was designed primarily for statistical machine learning of acoustic- and language-models for automatic spontaneous speech recognition, but it was also designed for the study of language variation.

So far, we have analyzed variations at different levels of language structures including, vowel devoicing, pitch-accent location in adjectives, coalescence of particle succession, moraic nasalization of particles, diffusion of the new potential verb forms, choice of phrase-final boundary pitch movements (BPM), and strength of the prosodic boundary preceding accented particle. In addition to these, analysis of word-form variation was conducted. The last analysis was concerned not only with individual lexical items, but also with the lexicon as a whole.

Lastly, I will touch briefly upon NIJL's new project that aims at the construction of a large-scale balanced corpus of cotemporary written Japanese, and its potential contribution to the study of spoken language.

KEYWORDS

Spontaneous speech, Corpus, Variation, Intonation, Japanese

1. INTRODUCTION

Linguistic variation is a statistical phenomenon. Prediction of the occurrence of a given variant is possible only on probabilistic ground even when the context in which the variant is located is given. The natural consequence is that the study of linguistic variation tends to requires large amount of data. This is particularly true when one wants to study variants whose occurrences are influenced by many factors encompassing both linguistic and social aspects of the target language.

Accordingly, those linguists who study variation are among the people who most enthusiastically welcome the availability of large-scale corpora and the rise of a new discipline called corpus linguistics. Today, as a matter of fact, substantial parts of introductory textbooks of corpus linguistics are devoted to the study of variation.

While this may be true for written language, the situation is drastically different for spoken language. It is also true with spoken language that the study of variation requires large amounts of data. What is drastically different from the case of written language is that there isn't any corpus of spoken language that could be used for the study of language variation.

This may appear odd given the facts that 50% of the data collected by the *Survey of English Usage* project (known today as the London-Lund Corpus), and 10% of the *British National Corpus* are devoted for spoken language. In the case of the SEU, the total amount of data, 500 thousand words, is too small to conduct complex analyses, and in the case of BNC, the transcription is too broad to get fine information about the phonetic details. In addition, probably the most important drawback is that both corpora do not provide speech sound files (In the case of BNC, we can listen to the speech materials in the British Library, but the materials are not publicly available).

In the field of speech engineering, on the other hand, statistical approach, hence corpus-based approach, has been the main-stream for at least 20 years. Spoken language corpora have been widely used for the purpose of automatic learning of language- and/or acoustic-models for automatic speech recognition (ASR) system. Later on, in 1990s, corpus-based speech synthesis was also developed. It turned out that it was possible to synthesize naturally-sounding speech just by making optimal concatenation of labeled speech sounds in the corpora.

The corpora compiled for speech processing purposes are, however, of little use for the study of language variation, because the speech material is hardly spontaneous. Typically, the materials in the corpora are spoken version of written texts like newspaper articles or so-called phonemically balanced sentences. These materials were pronounced, typically, by professional narrators to have as small a number of fluctuations as possible.

Accordingly, scholars of language variations had to compile databases, or corpora, of their own each time they started examining a new variable. This is needless to say a time-consuming effort. Moreover, those corpora constructed for personal use become rarely available for other researchers.

I didn't hesitate much, in February 1998, when the director general Seiichi Yamamoto of the ATR spoken language translation laboratory (currently professor of Doshisha University) called me and asked if I was willing to be one of two sub-leaders of a new speech processing project in which I was expected to design and compile a large corpus of spontaneous Japanese under the supervision of professor Sadaoki Furui of the Tokyo Institute of Technology. I could almost intuitively understand that it was a good occasion for the discipline of the study of language variation.

Almost a year later, we submitted a research proposal to the former Science and Technology Agency (currently a part of the Ministry of Education), and the proposal was accepted without much ado. This is how the *Spontaneous Speech: Corpus and Processing Technology* project got started in the spring of 1999. This was a five-year (1999-2003) joint project of National Institute for Japanese Language (NIJL), National Institute for Information and Communications Technology (aka Communications Research Laboratory till 2001), and Tokyo Institute of Technology.

Although the goal of the project was to develop a prototype system for the next generation ASR system that could recognize spontaneous speech, there was a clear consensus among the project members that development of large-scale spontaneous speech corpus was the key issue. In the first six months or so of the project, my colleagues and I concentrated our efforts in designing a corpus that could capture as much information as possible about the variability ---both physical and linguistic--- of spontaneous speech, based upon the belief that an optimal corpus of spontaneous speech designed for ASR system could be an excellent resource for the study of language variation as well.

2. OUTLINE OF THE CORPUS OF SPONTANEOUS JAPANESE

The spontaneous speech developed by the abovementioned project is known as the *Corpus of Spontaneous Japanese*, or CSJ. It was released in June 2004, and more than 300 copies have been purchased by researchers in various research institutions including universities, national laboratories, and companies.

2.1 THE SIZE

Table 1 shows the whole size of CSJ with respect to the numbers of words, speakers, talks, and total hour of recorded speech. The number of speakers is smaller than the number of talks because there were many speakers who provided more than one talks.

Table 1. Size of CSJ

N of running words	7,525,125
N of different speakers	1,417
N of talks	3,302
Total hour of speech	662

Table 2 shows the type of talks recorded in the CSJ. Parenthesized numbers of speakers were counted more than twice. As shown in the 'MODE' column of the table most of speech materials are devoted to monologues, but at the same time, they cover wide range of speaking styles ranging from read speech to free conversation. 'Reread speech' is the reading aloud of the transcription of spontaneous speech previously uttered by the same speakers.

Difference of talk types is an important factor of data analyses when we conduct linguistic analysis of language variation. Figure 1 shows the ratio (%) of word-form variation, i.e. the total number of non-standard variants divided by the number of total occurrence of the word in question multiplied by 100, as a function of the type of talks. There is a clear correlation between the ratio of word-form variation and the expected ranking of speaking style. See also section 4.2.

Table 2. Type of talks in CSJ.

TYPE OF TALKS	MODE	N FILE	N SPKER	HOUR
Academic Presentation Speech (APS)	Monologue	987	819	274.4
Simulated Public Speaking (SPS)	Monologue	1,715	594	329.9
Public Lectures (PL)	Monologue	19	16	24.1
Interview on APS	Dialogue	10	(10)	2.1
Interview on SPS	Dialogue	16	(16)	3.4
Task-oriented dialogue	Dialogue	16	(16)	3.1
Free dialogue	Dialogue	16	(16)	3.6
Reread speech	Monologue	16	(16)	5.5
Read speech	Monologue	507	(248)	15.5

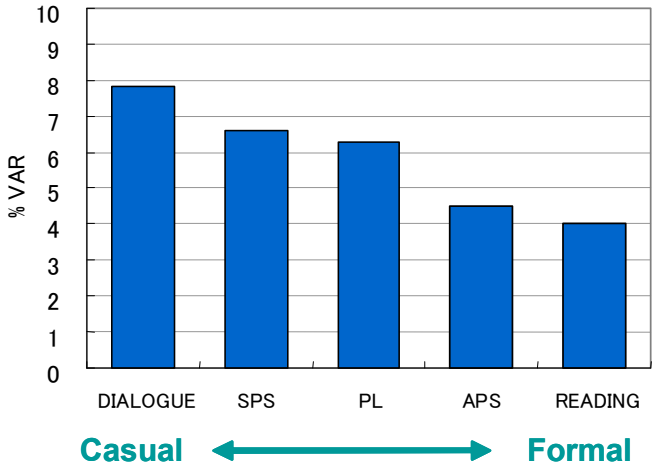

Figure 1. Correlation between the type of talks and the ratio of word-form variation.

2.2 SPEAKERS

Another important social factor of language variation is the age of speakers. Figure 2 shows the distribution of the CSJ speakers with respect to their birth years that were sectionalized for

every decade. There is clear difference between the speakers of APS and those of SPS.

APS speakers are heavily concentrated in their twenties, because most of the APS speakers were graduate students. SPS speakers, on the other hand, shows less skewed distribution compared to APS. This is because SPS speakers were recruited so that their age, and sex, show distribution as uniform as possible. Note, in passing, that the number of speakers shown in figure 2 is the total (cumulative) number of speakers (i.e. one and the same speaker may be counted more than twice when he/she gave more than two talks). If we count the number of different speakers as in figure 3, distribution of the SPS speakers is not as uniform as in figure 2, but it is still much more uniform compared to the distribution of the different APS speakers.

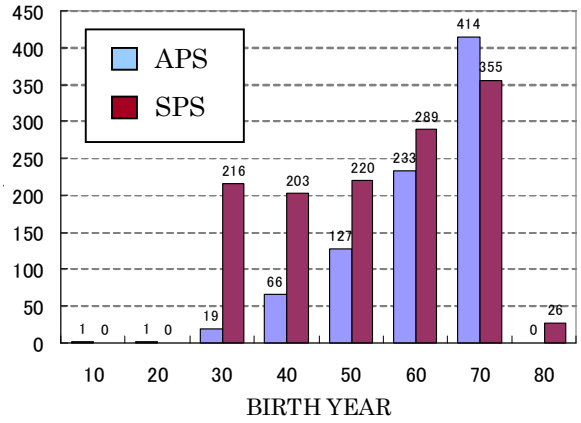

Figure 2. Number of total speakers as a function of their birth year.

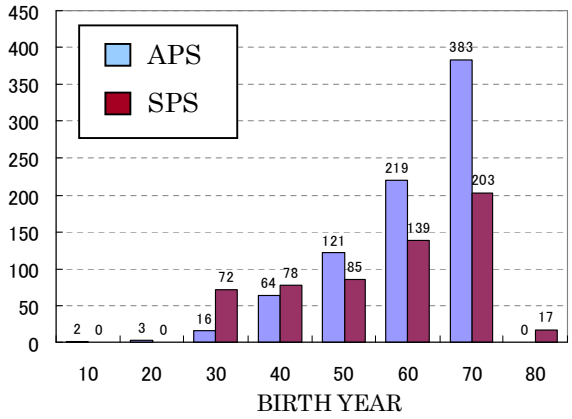

Figure 3. Number of different speakers as a function of their birth year.

Table 3. Distribution of the cumulative number of speakers.

SEX	APS	SPS	PL	READ	REREAD	INTERVIEW	Total
FEMALE	173	910	9	252	8	29	1381
MALE	814	805	10	255	8	29	1921
Total	987	1715	19	507	16	58	3302

Table 4. Distribution of the number of different speakers.

SEX	APS	SPS	PL	READ	REREAD & INTERVIEW	Total
FEMALE	138	331	6	(122)	(8)	470
MALE	681	263	10	(124)	(8)	947
Total	819	594	16	(246)	(16)	1417

Tables 3 and 4 show the distribution of speakers' sex as a function of talk types. Parenthesized numbers in the latter table indicate that the speakers were already counted as the speakers in other types of talks. All speakers of read speech, reread speech, and interview speech

were counted more than twice. Note that the speakers of reread speech and interviewees of interview speech belong to the same group of speakers.

2.3 ANNOTATIONS

As shown in figure 4, CSJ consists of several layers differing in the richness of annotation. This multi-layer structure was introduced into the corpus to satisfy incompatible needs of the corpus: richness of annotation and the size of corpus.

The Core of the CSJ includes half a million words and is the part of the corpus to which the cost of annotation was concentrated, the most crucial difference being the application of segmental and intonation labeling. Moreover, there are two more layers inside the Core that differ in the richness of annotation. The richest part in the Core are annotated with respect to segmental label, intonation label, dependency structure label, impression rating, and, topic structure label, in addition to the following annotations that are provided for all speech files, i.e., two-way transcription, two-say POS information, clause boundary information, impression rating, and, information about the speaker and the talk per se.

The rest of this section is devoted to a brief introduction to some of the CSJ annotations that will be referred to in the sections on language variation.

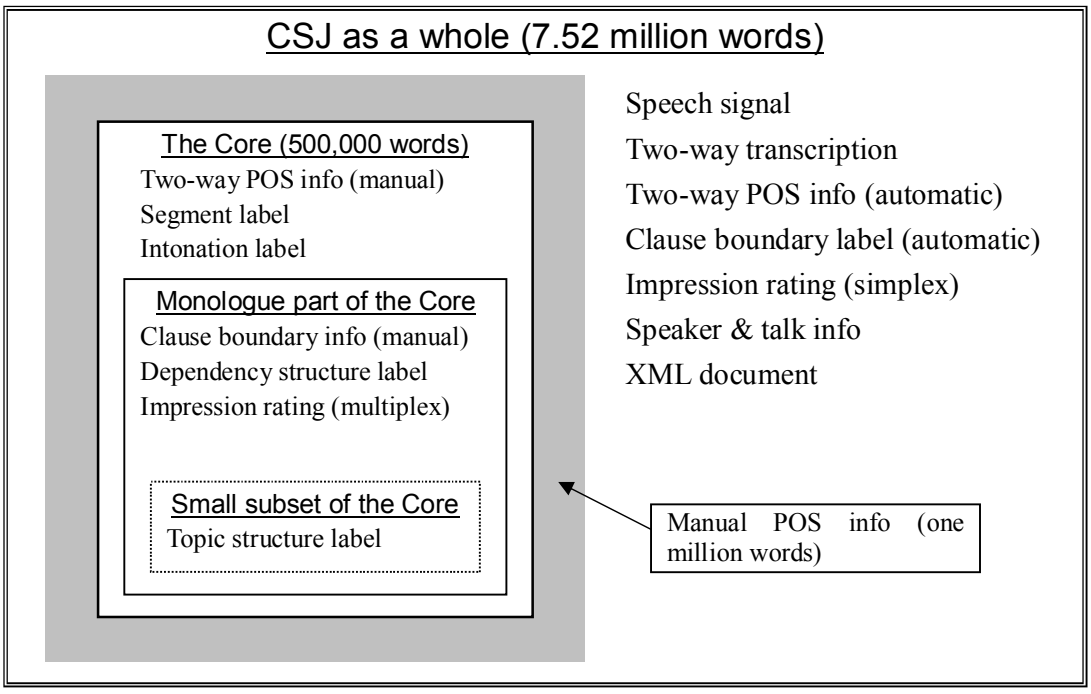

Figure 4. Layered structure of annotation in the CSJ.

2.3.1 Two-way transcription

Transcription of Japanese speech requires special treatment, because the language's orthography has a very high degree of freedom. There are, almost always, more than two ways of writing

down the same linguistic message. For example, there are at least four common ways of writing a compound verb /hanasi-au/ ("*discuss*"), viz., 話し合う、話合う、話しあう, and はなしあう. This flexibility in orthography could be a strong obstacle for the corpus search, needless to say.

CSJ overcame this problem by providing two independent transcriptions called orthographic and phonetic transcriptions. In the orthographic transcription, utterances were transcribed by Kanji (Chinese logographs) and Kana (Japanese syllabary) characters following the rules of orthography that we established for CSJ. The new orthography was designed so that there is no degree of freedom.

Phonetic transcription, on the other hand, uses Kana exclusively to transcribe the phonetic details of the utterances as exactly as possible within the limitation of a syllabary.

The combination of orthographic and phonetic transcriptions provides powerful tool for the search of word-form variations. Figure 5 shows schematically how the transcriptions could be used for a search. In this case, word-form variation of adverb morpheme {yahari} ("*after all*") was examined. The left hand string of Kanji and Kana 矢張り is the orthographic transcription of (the dictionary form of) the morpheme, while the right hand strings of ヤハリ, ヤッパリ,ヤッパ, ヤパ, ヤッパシ, etc. are the variants of the adverb as they are represented in the phonetic transcription. By making comparison of two transcriptions in this way, it is possible to extract useful information about word-form variations.

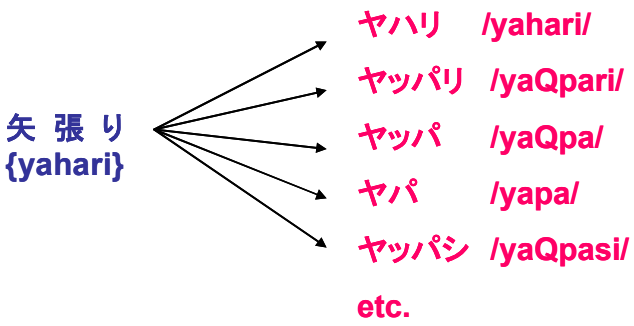

Figure 5. Orthographic and phonetic transcriptions.

2.3.2 Two-way POS information

CSJ provides two-way POS information of SUW (short unit word) and LUW (long unit word). The two-way POS analysis is required because, for one thing, Japanese is a so-called agglutinative language, and, for another, the definition of a word is heavily theory dependent in Japanese.

For example, the name of the institution to which the present author belong, {kokuricukokugokenkyuHzyo} consists of a single LUW noun, but at the same time it consists of three SUW nouns, {kokuritsu} ("*national*"), {kokugo}("*Japanese language*"), and, {kenkyuu} ("*research*"), which is followed by an affix morpheme {zyo} ("*institution*").

In this particular case, all nouns and affix are Sino-Japanese; there are of course cases where LUW is formed by concatenating Japanese native words and morphemes. For example, a single

LUW native verb of {mociagecuzukeru} (“*keep on lifting up*”) consists of two native SUW verbs {mociageru } (“*lift up*”), and aspectual native verb {cuzukeru} (“*keep on*”).

Note, in passing, that a SUW can consist of two constituents that can be treated as independent SUW. In the case of the above example, the SUW {mociageru} consists of two constituent verbs {mocu} (“*have*”)and {ageru} (“*lift*”); each of these can be used as a simplex verb SUW. In such cases, the constituent SUW should be simplex SUW (i.e., SUW that does not involve two SUW).

2.3.3 Intonation labels

X_JToBI (Maekawa et al. 2002, NIJL 2006), an extended version of the J_ToBI intonation labeling scheme (Venditti, 1997, 2005) was developed for the labeling of spontaneous speech. The new scheme is capable of expressing the details of the prosodic characteristics of spontaneous speech including, among other thing, boundary pitch movements (i.e., the characteristic movement of pitch at the phrase boundary) and their variation.

2.3.4 Impression rating

Impression rating is the subjective rating of various impressions that listeners perceive from spontaneous talks. CSJ provides two kind of impression rating data.

Each monologue talk (APS, SPS, and PL) was evaluated at the time of recording by a rater about speaking rate, speaking style, spontaneity of the talk, etc. Although different talks were evaluated by different raters (i.e., the rater was not uniform for all talks), the resulting impression rating data turned out to be very useful for the analysis of language variations. See 3.3 and 3.4 below. This data is called simplex impression rating data.

There is another type of impression rating data called multiplex data. Monologue talks in the Core were evaluated by 10 raters using psychological scales developed specially for CSJ (Yamazumi et al., 2006).

2.3.5 Clause boundary label

It is often very difficult to label the sentence boundary of spontaneous speech. It is in most cases possible, however, to label the boundary of syntactic *clauses*, i.e., the syntactic unit consisting of predicates and their complements. All transcription files of the CSJ were automatically classified with respect to the morphological characteristics of the predicates using the result of POS analysis (mostly SUW information). For all talks included in the Core, the results of the automatic classification were checked and, if necessary, corrected by human labelers.

3. ANALYSIS OF SOME SELECTED LANGUAGE VARIATIONS

In this section, results of some pilot studies about language variation will be presented. Examples are selected so that they cover as wide range of linguistic structure as possible. All examples use CSJ as the data source, needless to say. Note some of the studies were conducted while the

compilation of the CSJ was underway, in order to evaluate the usefulness of the corpus (Maekawa, 2004, Maekawa et al., 2003 for example). As the result, some of the results reported below did not use the current version of CSJ as its source.

3.1 VOWEL DEVOICING

It is well known that in Japanese close vowels, --/i/ and /u/--, are devoiced when they are preceded and followed by voiceless consonants. This is the typical environment of vowel devoicing in Japanese. Maekawa and Kikuchi (2005) analyzed a subset of CSJ-Core containing 427,973 vowels and reported several interesting findings.

Table 5. Rate of vowel devoicing as a function of the voicing of adjacent consonants.
C1=Preceding consonant, C2=Following consonant,
Co=Voiceless consonant, Cv=Voiced consonant.

VOWEL	C1	C2	VOICED	DEVOICED	%DEVOICED
a	Co	Co	12,214	262	2.10
	Co	Cv	18,570	92	0.49
	Cv	Co	24,943	481	1.89
	Cv	Cv	19,867	29	0.15
e	Co	Co	5,550	190	3.31
	Co	Cv	10,890	116	1.05
	Cv	Co	11,552	323	2.72
	Cv	Cv	11,388	29	0.25
i	Co	Co	1,475	12,124	89.15
	Co	Cv	10,556	2,219	17.37
	Cv	Co	9,200	126	1.35
	Cv	Cv	12,072	133	1.09
o	Co	Co	12,247	437	3.45
	Co	Cv	19,752	365	1.81
	Cv	Co	14,650	13	0.09
	Cv	Cv	16,802	14	0.08
u	Co	Co	1,732	9,267	84.25
	Co	Cv	11,851	3,133	20.91
	Cv	Co	5,562	127	2.23
	Cv	Cv	7,748	61	0.78

Table 5 compares the devoicing rate of five Japanese vowels under the four phonological environments defined in terms of the voicing of adjacent consonants. It shows that close vowels are not completely devoiced even under the typical environment of devoicing (i.e., C1 and C2 are both Co); it also shows that there isn't any environment where devoicing is completely avoided.

It has been pointed out by phoneticians that close vowels tend not to be devoiced in the

environment where more than two morae (syllables) could be sequentially devoiced. Words like /susi/ ("*sushi*"), /kucusita/ ("*sox*"), /kikuci/ (Japanese surname) and /fukusiki/ ("*duplex*") contain environments of sequential devoicing.

This tendency has been acknowledged by many phoneticians, but the mechanism of the avoidance, i.e., the mechanism that determines which vowel is to be devoiced and which is not, was not clearly recognized. Analysis of spontaneous speech revealed interesting tendency with respect to sequential devoicing.

Figure 6 shows the devoicing rate of two adjacent close vowels in the environment of sequential devoicing as a function of the combination of the manners of articulation of the mora-initial consonants. The notation like 'F/A' means that the consonant of the first close volwe ('V1') is fricative and that of the second close vowel ('V2') is affricate. There is a clear trading relationship between the devoicing rates of V1 and V2, with the sole exception of 'S/S'.

Speakers tended to avoid devoicing of V1 especially in the environment of 'F/F', 'S/S', and 'A/F'. In these environments, devoicing of V1 gives rise to two consecutive fricatives (including the last half of affricates) or consecutive stops. The consonant sequences like [kk] (as in /kikuci/), [kts] (as in the first half of /kucusita/, where /c/ is affricate) are phonetically realized as the consecutive occurrence of two sound spikes on the time dimension, and is often difficult to be perceived. Similarly, fricative sequences like [sʃ] (as in /susi/) or [tsʃ] (as in /kucusita/) can be difficult to be perceived. On the other hand, environments like 'F/A' and 'F/S' are easy to be perceived even when V1 is devoiced, because the consonant sequences are clearly punctuated by the presence of stops (including the first half of affricates).

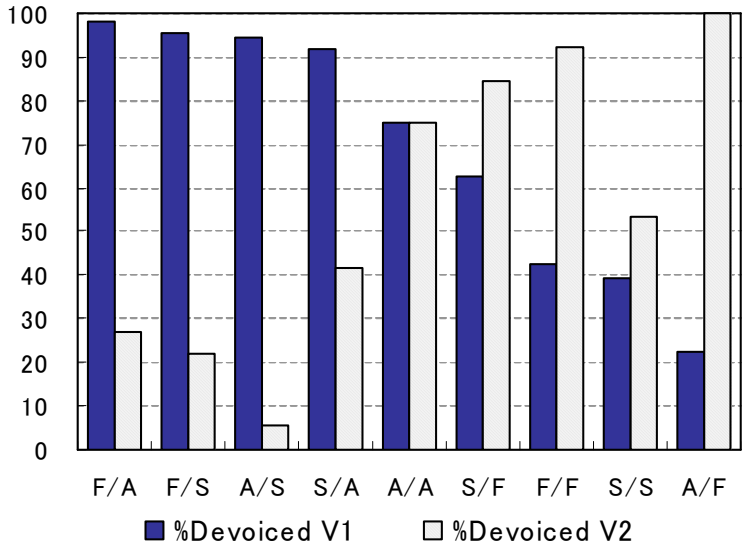

Figure 6. Devoicing rate of adjacent close vowels in the environment of sequential devoicing. V1: First vowel, V2: Second Vowel, F: Fricative, A: Affricate, S: Stop.

3.2 PHRASING OF ACCENTED PARTICLES

Most dialects of Japanese have lexically specified pitch accent. Tokyo Japanese that is the main target of the CSJ is no exception. Lexical items in these dialects are specified for the presence and absence of lexical accent, and in the case of presence, the location of pitch accent as well.

When lexical items are joined together to form an utterance, however, not all accents are realized as they are specified in the lexicon. There are rules of compound word accentuation, and there are also rules of accentual phrasing.

Accentual phrase (AP) is the most important unit of Japanese prosody in which at most one lexical accent can be specified (i.e., an AP is either accented or unaccented). Therefore, rules, or principles, of accentual phrasing has to determine which accent is to be deleted when there are more than two accents in the string of lexical items that are to be integrated into a single AP.

Most of the existing literatures on AP in Tokyo Japanese says that accent in the accented particles will be lost when they follow accented nouns, or verb sometimes, to form an AP. For example, particle /ma*de/ ("*to*"), where asterisk is used to denote lexical accent, will become unaccented when it follows accented nouns like /kyo*Hto+made/ ("*to Kyoto*") or /yo*ru+made/ ("*to the night*"), while it retains its accent when it follows unaccented nouns like /yokohama+ ma*de/ ("*to Yokohama*") or /yuHgata+ma*de/ ("*to the evening*").

This description is widely acknowledged. But astute observers of spontaneous Japanese are aware that this is not always the case. Figure 7 is taken from Maekawa and Igarashi (2006) that examined the behavior of two-mora accented particles that formed an AP with the immediately preceding accented lexical items in the CSJ.

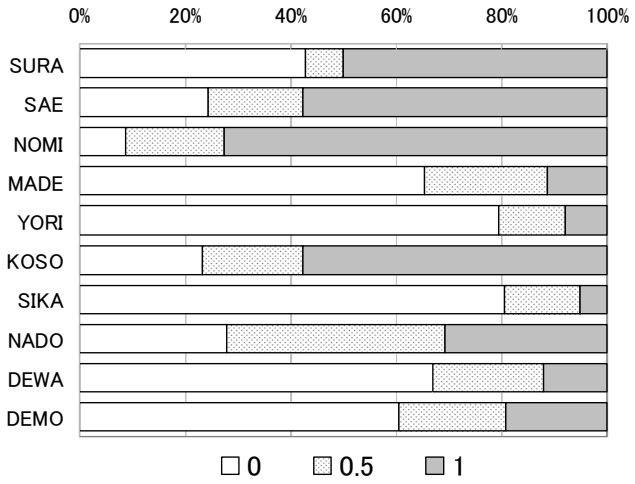

Figure 7. Prosodic independence of two-mora accented particle.
0: particle accent is deleted, 1: accent not deleted, 0.5: two labelers did not agree with respect to particle accent.

In this figure, accentedness of 10 two-mora accented particles was compared. The shaded bar (shown as "1" in the legend) represents the percentage of cases where particle accent was not deleted, and, the dotted bar represents the case where two raters gave different judgment about

the accentedness of particle. The open bar, accordingly, represents the cases where accent in the particles were deleted. This figure suggests strongly that the rule about AP formation of accented particles is virtually an optional rule, or there might be many hitherto unknown factors that prevent the rule from being applied.

Maekawa and Igarashi (2006) examined the effects of various linguistic and extra-linguistic factors on the phrasing and concluded that the most influential factor was the semantic property of particles. Particles whose semantic function is emphasis and/or limitation tend to constitute an AP of their own.

3.3 WORD COALESCENCE

Under some circumstances, function words like particles and auxiliary verbs can be merged with their adjacent words. This phenomenon is called word coalescence. Among the most frequent word coalescences of Tokyo Japanese, coalescence of /de/ and /wa/ into /zya/ was analyzed. This coalescence can be found in two word sequences that are completely distinct from a linguistic point of view; case particle /de/ followed by topic particle /wa/ on the one hand, and, auxiliary verb /da/ in its adverbal form (i.e., /de/) followed by the topic particle, on the other.

Figure 8 shows the result of a decision-tree analysis of the coalescence. As shown in the top box, the overall coalescence rate is 22.0%. But we can predict the occurrence of coalescence more accurately if we know the POS of /de/; the rate becomes 1.7% and 42.6% when /de/ is particle and auxiliary verb respectively. In the save vein, but to much lesser extent, factors like type of talks (APS or SPS), speaking style (formality), and spontaneity (spontaneous of prepared) could be useful for the prediction of coalescence. Note the last two factors mentioned above are part of simplex impression rating data. Note also that all these factors, both linguistic and extra-linguistic are provided in the CSJ.

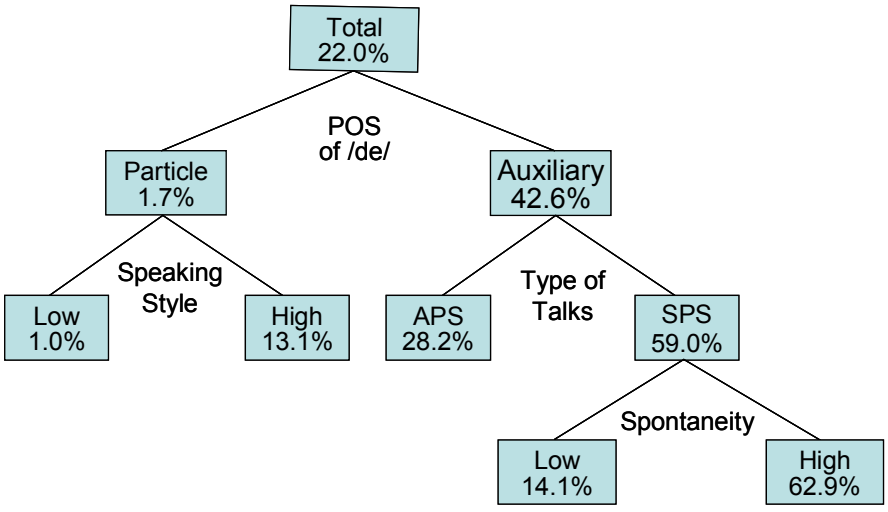

Figure 8. Decision-tree of word coalescence /de/+/wa/ => /zya/.
Digits show the rate of coalescence.

3.4 BOUNDARY PITCH MOVEMENTS

Boundary pitch movement (BPM) is those characteristic intonations that mark the end of accentual phrases. This is one of the areas of Japanese intonation study that is very interesting but underdeveloped.

In the X-JToBI labeling, the normal falling tune was marked by the label 'L%' and all BPM were labeled as one of the followings: 'L%H%' (rise), 'L%LH%' (another type of rise, called "insisting rise"), 'L%HL%' (rising-falling tune), 'L%HLH%' (rising-falling-rising tune). In addition to these basic categories, intonation variations were represented by additional labels like 'FR' (standing for "floating rise", a variant of L%H% and L%HL%) and 'PNLP' ("penult non-lexical prominence", a variant of L%HL%).

Figure 9 shows the occurrence rates of L%H% and L%HL% as a function of the impression rating of speaking style and spontaneity (Maekawa et al., 2003). It is interesting to see that the behaviors of the two BPM are complementary. The rate of rising tune (L%H%) correlates positively and negatively with speaking style and spontaneity respectively, while the rate of rising-falling tune (L%HL%) correlates negatively and positively with speaking style and spontaneity. Note that higher number in speaking style means that the speaking style is more formal.

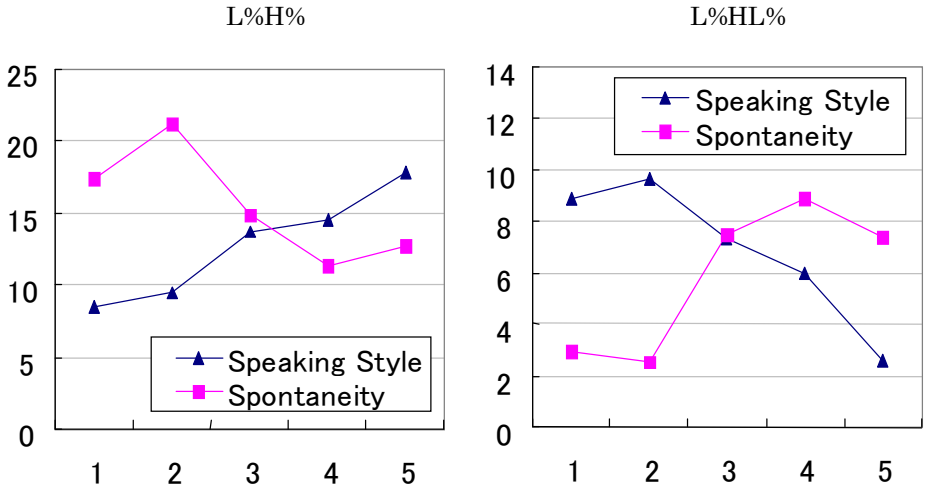

Figure 9. Relationship between the occurrence rates of BPMs [%] and the impression rating of speaking style and spontaneity (abscissa).

As mentioned earlier, X-JToBI labeling has the inventory of 'PNLP' that marks special variant of L%HL%. In ordinary L%HL% intonation, the intonation peak is located in the last syllable of accentual phrase. There are, however, cases where the peak locates in the penult syllable. This is the case to which 'PNLP' label is applied.

Figure 10 shows the relationship between the rate of PNLP to the total occurrence of L%HL% as a function of talk types (either APS or SPS), speaking style, and, spontaneity. It is interesting to see that the ratio of PNLP variants correlates positively and negatively with

speaking style and spontaneity; this is the same pattern of correlation as in the L%H% BPM.

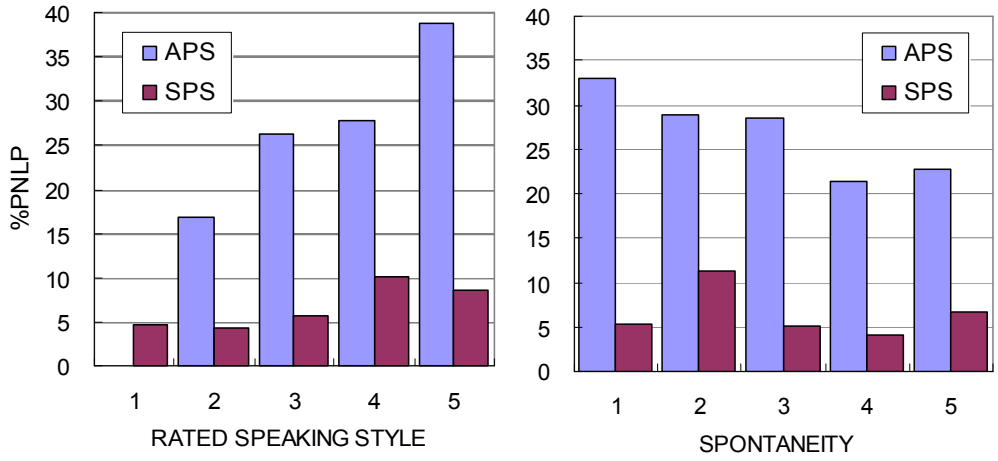

Figure 10. Correlation between the occurrence rate of PNLP variant of L%HL% and impression ratings of speaking style and spontaneity.

3.5 POTENTIAL VERB (INTROSPECTION AND BEHAVIOR)

Variation of potential verbs is one of the most well-known variations in the verb-morphology of the present-day Japanese. Traditionally, potential forms of vowel-ending verbs like {miru} (‘*see*’), and {taberu} (‘*eat*’) are derived by inserting a potential suffix {rare} between their roots and suffix (i.e., /ru/), the resulting forms being /mi-rare-ru/ and /tabe-rare-ru/. During the past hundred years or so, however, new potential suffix /re/ has been emerging steadily. This is per se an interesting morphological variation. But the analysis of potential verbs provides very interesting finding about the survey methodology in the study of variation (Maekawa, 2005b).

Figure 11 is the result of questionnaire survey about the potential form of {kuru} (‘*come*’) done by Japanese Government’s Agency of Cultural Affairs in 2001. In this survey, the subjects were shown the list of traditional /ko-rare-ru/ and innovative /ko-re-ru/ (both mean ‘*able to come*’), and asked which one they used. In this figure, innovative form overtook traditional form in the group of subject born in the years 1971-80.

On the other hand, Figure 12 is the result obtained by analyzing CSJ. In this figure traditional form was overtaken by innovative form as early as in the group of subjects born in 1940-49. So, there is at least about 30 year difference between the two surveys with respect to the timing of innovative form’s take-over.

The most straightforward interpretation of this discrepancy would be that most subjects of the questionnaire survey were influenced by their norm of writing, probably without knowing it. Use of innovative forms in writings is still exceptional even among the subjects who use innovative forms constantly in their speech. Needless to say, the data in CSJ is the ‘real’ recording of the subjects’ speech behavior without being biased by speakers’ incorrect introspection on their own speech behavior. Data of CSJ can be used to check the validity of questionnaire survey in this way.

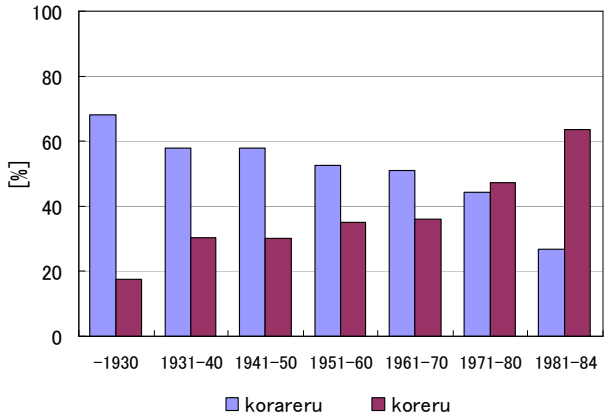

Figure 11. Result of a questionnaire survey about the use of potential form of {kuru} as a function of speakers' birth year.

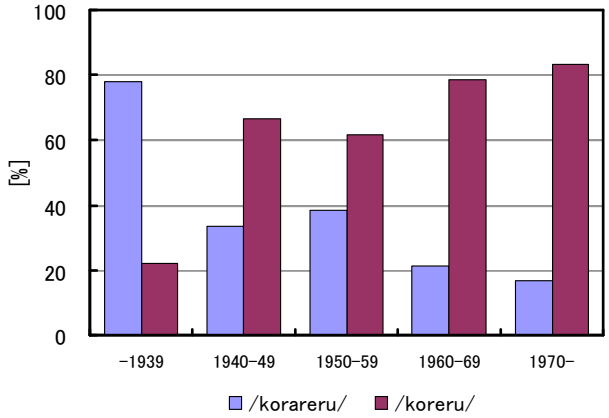

Figure 12. CSJ data about the use of potential form of {kuru} as a function of speakers' birth year.

4. ANALYSIS OF WORD-FORM VARIATION

So far, we have seen results of pilot surveys about language variations in the levels of phonology, morphology, and prosody. In the rest of this paper, I will present the result of ongoing study concerning the variation of word-forms. This is not the analysis of particular lexical items, but the overall survey of the lexicon as a whole (Maekawa, 2005a, 2005b).

4.1 DATA

In this study, two types of data about word-from variation were extracted from the CSJ: they are tentatively called phonetic and morphological variations.

Phonetic variation, or P-variation, was recorded in the phonetic transcription by using the tag (W) as in the following examples; (W kokoH; kokoro), (W deHtabeHsu; deHtaHbeHsu), and (W kakemaHru; kakemawaru). In these examples, observed word-forms were recorded as the first element of the tag. The second element of the tag, separated from the first one by a

semicolon, is the 'standard' word-from.

First example deals with the case where consonant /r/ is dropped and replaced by a long vowel (/H/ represents the second mora of a long vowel). The second example is concerned with shortening of lexically specified long vowel. And, the last one is concerned with the drop of /w/ and replacement by a long vowel. All these examples are concerned with articulatory weakening.

The examples shown above occurred very frequently in spontaneous Japanese, but the tag is also applied for sporadic variations. For example, among the 8240 occurrences of lexeme {niQpon}, the country name of Japan, /nihoN/ and /niQpoN/ occurred 7977 and 195 times respectively. In addition to these two, there were other sporadic variants like /nioN/ (39 times), /nihoHN/ (16), /zihoN/ (2), /ioN/ (2) etc.

Morphological variation, or M-variation, is the word-form variation that is not labeled by the tag (W). For example, none of the two variants of the country name of Japan, /nihoN/ and /niQpoN/ are marked by (W). Similarly, variants of the verb meaning to '*say*', --/iu/ and /yuH/--, and variants of the first-person singular pronoun, --/watashi/ and /atashi/-- are not marked by (W) altogether. The tag (W) is not applied to these variants for two reasons. For one, it is practically impossible to determine which variant is the 'standard' one. For another, some of these variants are not phonetically motivated hence inappropriate to be marked by (W).

Put differently, it was our principle to apply the (W) tag to the variations that are either sporadic or caused by articulatory weakening, or both. On the other hand, the tag is not applied to the cases where most native speakers are aware of the existence of the variation. In fact, the examples of M-variants shown above are usually found among the direction words in ordinary Japanese dictionaries. This is the direct consequence of speakers' awareness about the variation and variants.

Due to the limitation of pages, detailed explanation about how M-variations were extracted has to be omitted with the exception of the following two important points. First, when we talk about 'word-form' rather than 'word', every conjugation form of a conjugational word will be counted as different word-forms. Ending-form, adnominal-form, hypothetical form etc. of a verb, for example, will be recognized as separate word-forms. Second, we hypothesized that every word-form has only one 'standard' form, which is called dictionary form or DF. As will be discussed in 4.5, this is clearly too strong a hypothesis, but this is required for the automatic extraction of M-variations. The process of M-variation extraction is described in Maekawa (2005).

The data that will be analyzed below contains 302,019 M-variations. Because there were 130,951 P-variations in the corpus, the total number of variations was 432,970. Needless to say, these numbers represent the total (or 'running') number of word-forms. The number of different word-forms was 11,379 including both P- and M-variations.

4.2 CORRELATION WITH TALK TYPE

There is correlation between the total occurrence rate encompassing P- and M-variations and talk types. We have already seen this in figure 1 earlier. The general tendency is that the variation rate becomes higher in talks with lower formality and less spontaneity. It is important to note,

however, that even in the least spontaneous talk type of 'READING', about 4% of words showed word-form variation. This fact suggests strongly that the presence of word-form variation is virtually inseparable from our speech behavior.

4.3 WORD-FORMS WITH HIGH FREQUENCY OF VARIATION

Table 6 lists 20 word-forms that showed the highest frequency of non-DF variants. The fourth column is the total frequency of the word-forms in question. The fifth column is the frequency of variants other than the 'standard' form (DF). The sixth column is the ratio of fifth column over the fourth. And, the last column is the number of speakers who uttered the word-form at least once.

The most important fact concerning this table is that the sum of the frequency of non-DF variants (the 5th column) reached as many as 325,639 and covers about 75% of the total number of non-DF variants in the corpus.

Table 6. Twenty word-forms that showed the highest frequency of variations.

LEXEME	GLOSS	POS (CF)	N	Freq. Non-DF	% Non-DF	N of Speaker
{iu}	*'say'*	Verb (adnominal form)	132,818	132,332	99.6	1,411
{no}	*'of'*	Adnominal particle	153,521	79,829	52.0	1,326
{keredo}	*'but'*	Conjunction particle	47,032	26,534	56.4	1,092
{nani}	*'what'*	Pronoun	23,067	17,140	74.3	1,054
{iu}	*'say'*	Verb (ending form)	9,155	7,991	87.3	1,031
{Qte}	---	Adverbial particle	50,704	7,834	15.5	956
{niQpoN}	*'Japan'*	Noun	8,242	8,045	97.6	849
{kurai}	*'even'*	Adverbial particle	8,947	7,758	86.7	951
{ni}	*'at'*	Case particle	206,614	7,568	3.7	1,097
{yahari}	*'after all'*	Adverb	11,746	7,022	59.8	706
{sore}	*'that'*	Pronoun	44,000	6,016	13.7	767
{yoi}	*'good'*	Adjective (adnominal form)	5,950	5,177	87.0	934
{yoi}	*'good'*	Adjective (ending form)	4,446	4,026	90.6	866
{moH}	*'anymore'*	Adverb	18,501	3,669	19.8	674
{desu}	*Copula*	Aux. verb (ending form)	141,084	3,431	2.4	624
{de}	*'and then'*	Conjunction	55,717	3,290	5.9	756
{mina}	*'everyone'*	Noun	4,309	2,634	61.1	593
{mono}	*'thing'*	Noun	31,794	2,373	7.5	593
{watasi}	*'I'*	Pronoun	15,749	2,367	15.1	395
{soH}	*'so'*	Adverb	29,698	2,327	7.8	619

The rate of variation shown in the sixth column of table 6 is not necessarily the ratio of a single variant. Rather, it was usually the case that multiple variants were observed for a single lexeme. Table 7 is prepared to examine this problem. The second column is the number of different variants observed more than twice in the current data. As can be seen from the table, some

word-forms have more than 50 different variants. The third column of the table shows the coverage by the most frequent variant, i.e., the frequency of the top variant divided by the number shown in the fifth column of table 6. Similarly, the fourth column shows the cumulative coverage by the top 3 variants. In 14 word-forms out of 20, top 3 variants cover more than 99% of the variants, and, there are only two items whose cumulative coverage does not reach 95%, {yahari} and {sore}. This table shows convincingly that it is not necessary to make a long list of non-DF variant to cover the majority of total variation.

Table 7. Coverage of non-DF variants by top variants (Same order of row as in table 6).

LEXEME (CF)	N of Different Variants	Coverage by the Top Variant (%)	Coverage by the Top 3 Variants (%)	Top 3 Variants (From left to right)
{iu} (adnom.)	31	90.3	99.6	/yuH/, /yu/, /yuu/
{no}	15	52.2	99.7	/N/, /no/, /do/
{keredo}	53	53.2	98.5	/kedo/, /keredo/, /keHdo/
{nani}	25	73.9	97.4	/naN/, /nani/, /naNni/
{iu} (ending)	11	90.3	99.0	/yuH/, /yu/, /yuu/
{Qte}	22	82.6	99.2	/Qte/, /te/, /Qti/
{niQpoN}	6	96.8	99.6	/nihoN/, /niQpoN/, /nion/
{kurai}	6	88.7	99.7	/gurai/, /kurai/, /gura/
{ni}	33	96.3	99.8	/ni/, /N/, /i/
{yahari}	56	49.3	91.9	/yaQpari/, /yahari/, /yaQpa/
{sore}	98	85.8	93.8	/sore/, /soe/, /soi/
{yoi} (adnom.)	5	86.0	99.7	/iH/, /yoi/, /i/
{yoi} (ending)	7	91.1	99.4	/iH/, /yoi/, /i/
{moH}	20	80.1	99.3	/moH/, /mo/, /mu/
{desu} (ending)	60	97.4	99.2	/desu/, /esu/, /su/
{de}	33	91.6	98.9	/de/, /Nde/, /te/
{mina}	6	63.3	99.3	/miNna/, /mina/, /miNHna/
{mono}	25	92.3	99.4	/mono/, /moN/, /moH/
{watasi}	34	83.5	98.0	/watasi/, /atasi/, /tasi/
{soH}	28	92.0	99.0	/soH/, /so/, /soQ/

4.4 WORD-FORMS WITH HIGH RATE OF NON-DF VARIATION

It is important to note that tables 6 and 7 are concerned with the absolute frequency of non-DF variants, and not with the rate of variation. Consequently, there are word-forms whose variation rate is not so high but listed in table 6 because its occurrence frequency is quite large. Particle /ni/ and copula /desu/ are good example.

Table 8 shows the top 10 lexemes of the highest occurrence rate of non-DF variants. There are 3 items ---{niQpon}, {iu}, and {yoi}--- shared by tables 6 and 7. Note, in passing, word-forms whose frequencies were fewer than 10 were removed from the computation for this table.

Table 8. Word-forms of the highest occurrence rates of non-DF variants

LEXEME	POS (CF)	N (including DF)	N of Different Variants	Freq Non-DF	% Non-DF
{iu}	Verb (adnominal form)	132,818	31	132,322	99.6
{meHN}	Noun	162	2	157	98.1
{niQpoN}	Noun	8,242	6	8,045	97.6
{kaNzuru}	Verb (adnominal form)	274	2	266	97.0
{simyureHsyoN}	Noun	227	5	226	96.9
{enueicikeH}	Noun	183	7	176	96.2
{taiiku}	Noun	151	3	145	96.0
{syoHzuru}	Verb (adnominal form)	116	2	106	94.0
{poi}	Suffix (adnominal form)	145	2	136	93.8
{yoi}	Adjective (ending form)	4,446	7	4,026	90.6

4.5 ENTROPY OF WORD-FORMS

In the computation of variation rates presented above, we hypothesized that there in one and only one 'standard' DF for a given word-form, but this is a problematic hypothesis. There are many word-forms that have more than two 'standard' forms. For example, lexeme {niQpon} has two frequent word-forms /nihoN/ and /niQpoN/ both of which are registered in dictionaries. Similar examples include {yoi} ("*good*", frequent DF being /yoi/ and /iH/), {iu}("*say*", /iu/ and /yuH/), {iku} ("*to go*", /iku/ and /yuku/), {watasi} (1st person pronoun, /watasi/ and /atasi/), {mina} ("*everybody*", /mina/ and /miNna/), and so forth.

In these word-forms, the rate of variation could change drastically depending on the choice of DF. In the case of {niQpoN} for example, the current rate of 97.6% (see table 6) becomes 2.4% or even less, if we adopt /nihoN/ as the DF. Note that this is not at all a strange choice for the native speakers of Japanese. Clearly, an index of variability that does not make reference to 'standard' word-form is needed to avoid this kind of indeterminism in the quantification of word-form variation.

Entropy (in the sense of information sciences) is one such index. Entropy H of a probabilistic event E is the index of the predictability of E and is defined as

$$H = \Sigma\, p_i\, I(p_i)$$

where p is the probability distribution of the event E and $I(p_i)$ is defined as

$$I(p_i) = -\log_2 p_i$$

$I(p_i)$ is called the information (or information quantity) of the event.

If H=1 (unit of entropy is BIT), the event is as predictable as the result of coin tossing; the entropy of dice is about 2.585, showing that the prediction of dice is much more difficult than coin tossing.

Table 9 shows the entropy of word-forms previously shown in tables 6 and 7. As predicted, entropy of {iu} in its adnominal form, or, {niQpoN} is low because most of the occurrences is occupied by a single word-form whish happened not to be identified as the DF. On the other

hand, entropy of {no}, {nani}, and {mina} are about 1.0 because in these items two equally frequent word-forms are observed. And, lastly, entropy of {yahari} is higher than 2.0 because there are many word-forms that are used more or less frequently: for the total occurrence of 11,746, /yaQpari/ (N=5793), /yahari/ (3999), /yaQpa/ (998), /yaQpai/(256), /yaQpasi/ (112), /pari/ (112) and so forth.

Table 9. Entropy (H) of word-forms shown in tables 6 and 7.

LEXEME	GLOSS	POS (CF)	N	Freq. Non-DF	% Non-DF	H
{iu}	*'say'*	Verb (adnominal form)	132,818	132,332	99.6	0.587
{no}	*'of'*	Adnominal particle	153,521	79,829	52.0	1.033
{keredo}	*'but'*	Conjunction particle	47,032	26,534	56.4	1.477
{nani}	*'what'*	Pronoun	23,067	17,140	74.3	1.106
{iu}	*'say'*	Verb (ending form)	9,155	7,991	87.3	0.628
{Qte}	---	Adverbial particle	50,704	7,834	15.5	0.899
{niQpoN}	*'Japan'*	Noun	8,242	8,045	97.6	0.251
{kurai}	*'even'*	Adverbial particle	8,947	7,758	86.7	0.546
{ni}	*'at'*	Case particle	206,614	7,568	3.7	0.262
{yahari}	*'after all'*	Adverb	11,746	7,022	59.8	2.010
{sore}	*'that'*	Pronoun	44,000	6,016	13.7	1.146
{yoi}	*'good'*	Adjective (adnominal form)	5,950	5,177	87.0	0.687
{yoi}	*'good'*	Adjective (ending form)	4,446	4,026	90.6	0.515
{moH}	*'anymore'*	Adverb	18,501	3,669	19.8	0.799
{desu}	*Copula*	Aux. verb (ending form)	141,084	3,431	2.4	0.251
{de}	*'and then'*	Conjunction	55,717	3,290	5.9	0.566
{mina}	*'everyone'*	Noun	4,309	2,634	61.1	1.076
{mono}	*'thing'*	Noun	31,794	2,373	7.5	0.464
{watasi}	*'I'*	Pronoun	15,749	2,367	15.1	1.650
{soH}	*'so'*	Adverb	29,698	2,327	7.8	0.516

5. CONCLUSION AND PROSPECT

As shown by these examples, CSJ is a powerful tool for the analysis of language variation. It is the current author's wish to conduct these kinds of surveys more systematically to grasp the entire picture of language variation in spoken Japanese. And, ultimately, the result obtained from the CSJ should be compared to the variation of written language to get the full picture of the variation in the Japanese language.

At this point, however, we encountered a critical obstacle: the lack of reliable balanced corpus of written Japanese. Currently available database of contemporary written Japanese includes archives of newspaper articles (containing probably more than half billion words), but nothing else.

To overcome this problem, the NIJL started in the spring of 2006 a new project aiming at the construction of written language corpus (aka BCCWJ, balanced corpus of contemporary

written Japanese) of about 50 million words covering books, magazines, newspapers, internet texts, and more. Luckily, we were successful in obtaining a new research national grant for this corpus project in July 2005 for the compilation of additional 50 million words. Figure 13 shows the outline of the structure of BCCWJ (Yamazaki, 2006, Maekawa 2006, in press).

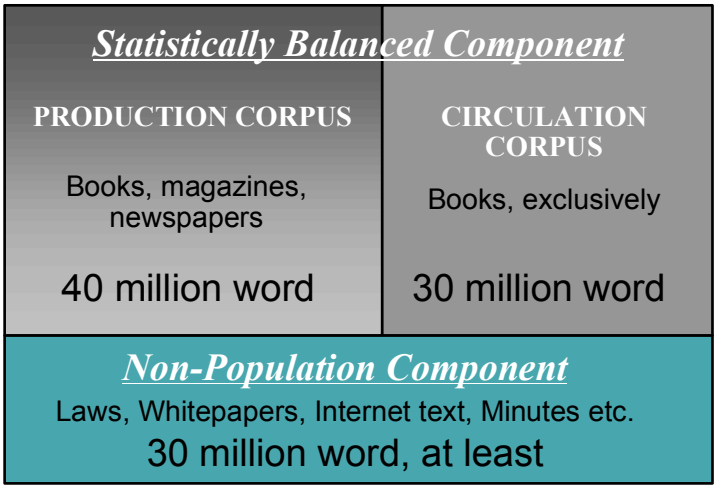

Figure 13. Outline of the BCCWJ.

As shown in the figure, there are three components in the corpus. The upper half of the figure is the "statistically balanced" component. Samples in this component are selected using the technique of random sampling. The statistically balanced component can be divided into two sub-corpora namely, "production corpus" and "circulation corpus".

These sub-corpora differ with respect to their sampling population. The population of the production corpus includes all books, magazines, and nation-wide newspapers published during the years 2001-2005. The size of this sub-corpus will be about 40 million words.

The population of the circulation corpus, which is also called "library corpus", is the whole body of books (excluding magazines and newspapers) to which ISBN (International Standard Book Number) are assigned and registered in at least one public library of the Tokyo metropolis. This population includes more than one million different book titles covering the years after 1980, when ISBN began to be adopted by Japanese publishers. The size of this sub-corpus will be about 30 million words.

The last component, located in the bottom of the figure, is called "practical" or "non-population" component. This component includes various sub-corpora that NIJL requires for its language planning study purposes, and those sub-corpora that cover the type of texts that are interesting, as well as important, for corpus linguistic studies, but are not covered sufficiently by the corpora in the statistically balanced component. They include samples of laws, governmental whitepapers, public relation magazines of local authorities, authorized textbooks, internet texts of BLOG and bulletin board, minutes of the Diet, bestsellers, and so forth. Currently, the compilation of BCCWJ is underway, and some of the design issues are still subject

to change.

I believe firmly that this corpus, when used with the CSJ, will provide a long-standing infrastructure for the study of the Japanese language including many applied fields like language pedagogy, language planning, Japanese information processing, and so on, not to mention the study of language variation.

ACKNOWLEDGEMNT

I'm very grateful for my colleagues and former colleagues at the NIJL who participated in the CSJ project and analysis of language variation; Hanae Koiso, Hideki Ogura, Masaya Yamaguchi, Hideaki Kikuchi, Takayuki Kagomiya, Wataru Tsukahara, Kiyoko Yoneyama, Masako Fujimoto, Kenya Nishikawa, Yoko Mabuchi, Yohichi Maki, Kenji Yamazumi, Takehiko Maruyama and Yosuke Igarashi.

REFERENCES

Maekawa, Kikuo (2004). "Design, Compilation, and Some Preliminary Analyses of the Corpus of Spontaneous Japanese", In K. Yoneyama and K. Maekawa eds. *Spontaneous Speech: Data and Analysis*, Tokyo:The National Institute for Japanese Language, pp.87-108.

Maekawa, Kikuo (2005a). "Toward a Pronunciation Dictionary of Japanese: Analysis of CSJ", *Proceedings of Symposium on Large-Scale Knowledge Resources (LKR2005)*, Tokyo Institute of Technology 21st Century COE Program, pp.43-48.

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

Maekawa, Kikuo (2006). "Kotonoha, the Corpus Development Project of the National Institute for Japanese Language." *Language Corpora: Their Compilation and Application (Proceedings of the 13th NIJL International Symposium)*, Tokyo, pp.55-62.

Maekawa, Kikuo (In press). "Language corpus development initiative in Japan", *LST Newsletter*.

Maekawa, Kikuo and Yosuke Igarashi (2006). "Prosodic independence of bimoraic accented particles: Analysis of the Corpus of Spontaneous Japanese", *Journal of the Phonetic Society of Japan*, 10 (2), pp.33-42 (In Japanese).

Maekawa, Kikuo and Hideaki Kikuchi (2005). "Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report". In J. van de Weijer, K. Nanjo, and T. Nishihara, eds., *Voicing in Japanese*, Mouton de Gruyter, pp.205-228.

Maekawa, Kikuo, Hideaki Kikuchi, Yosuke Igarashi and Jennifer Venditti (2002). "X-JToBI: An extended J_ToBI for spontaneous speech", *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP2002)*, Denver,Colorado,USA, 3, pp.1545-1548.

Maekawa, Kikuo, Hanae Koiso, Hideaki Kikuchi and Kiyoko Yoneyama (2003)."Use of a large-scale spontaneous speech corpus in the study of linguistic variation", *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, International Phonetic Association, pp.643-646.

National Institute for Japanese Language (2006). *Construction of the Corpus of Spontaneous Japanese*. (NIJL Research Report No. 124), National Institute for Japanese Language, Tokyo (In Japanese).

Venditti, Jennifer (1997). "Japanese ToBI labeling guidelines." In K. Ainsworth-Darnell, M. D'Imperio (eds.), *Papers from the Linguistics Laboratory, Ohio State University Working Papers in Linguistics*, 50, 127-162 [First distributed as web document in 1995].

Venditti, Jennifer (2005). "The J_ToBI model of Japanese intonation". In Jun, S. A. (ed.) *Prosodic typology: The phonology of intonation and phrasing*. New York: Oxford University Press.

Yamazaki, Makoto (2006). "Design of NIJL balanced corpus of contemporary written Japanese", *Language Corpora: Their Compilation and Application (Proceedings of the 13th NIJL International Symposium)*, Tokyo, pp.63-70 (In Japanese).

Yamazumi, Kenji, Takayuki Kagomiya, Yohichi Maki, and Kikuo Maekawa (2006). "Psychological scale for the impression rating of monologue". *The journal of the acoustical society of Japan*, 61 (6), pp. 303-311 (In Japanese).

URL

Some of the papers listed above are downloadable in the internet. See,

http://www2.kokken.go.jp/~kikuo/public/KMHP1.html (Japanese), and
http://www2.kokken.go.jp/~kikuo/public/KMHP1.html (English).

Visit also English homepage of CSJ;

http://www2.kokken.go.jp/~csj/public/index.html

If you read Japanese, you can obtain information about BCCWJ from the following two URL;

http://www2.kokken.go.jp/kotonoha/
http://www.tokuteicorpus.jp/