Quantitative Analysis of Word-form Variation Using a Spontaneous Speech Corpus

Kikuo Maekawa
National Institute for Japanese Language,
Tokyo Institute of Technology, and
Hitotsubashi University
kikuo@kokken.go.jp

1. INTRODUCTION

This paper examines the possibility of extracting valuable information for the amelioration of a pronunciation dictionary. While pronunciation dictionary refers to those dictionaries used by humans, most of the facts shown below would be useful for the design and construction of a pronunciation dictionary for automatic speech recognition. As a matter of fact, there are engineering papers that have analyzed the same corpus as analyzed in this paper. Readers who are interested in this field are asked to see Akita and Kawahara (2005).

Traditional pronunciation dictionary is essentially a list of possible surface word-forms of lexemes described in a *qualitative* manner. In addition to the list, some dictionaries provide frequency information for some word-forms in a qualitative manner like "*word-form A is more frequent that word-form B*". Also, sometimes, information about style difference is involved, like "*word-form A is more formal than B.*" It is well known, however, that *Longman Pronunciation Dictionary* (LPD, hereafter) has recently broken fresh ground by providing *quantitative* information about the way words were pronounced (Wells, 1990). By using this dictionary, it is possible to know, for example, that the last syllable of '*attitude*' is pronounced as /tu:d/ in 85% of cases, and the ratio of /tju:d/ remains as low as 15% in American English. Also, the dictionary provides many figures of age- and/or style-stratification of word-forms. The advancement achieved by LPD is clear and uncontradictable. But it doesn't mean that there is no room for further advancement.

One problem of the quantitative data recorded in the LPD could be found in the way the data was acquired, i.e. a questionnaire. Although questionnaires are an excellent, and under many circumstances the only method that enables simultaneous acquisition of coherent data over many words, their success depends heavily upon the informants' ability to retrospect one's own verbal behavior. When questionnaires are used for data acquisition of speech behavior, there could be a serious drawback as will be shown later (see 7.3), as informants' responses could be biased by their sense of linguistic norm, or, as it is sometimes the case that it is simply too difficult for them to retrospect their own pronunciation. Linguists and phoneticians have long believed that there was no effective way of observing many people's real speech behavior over many lexemes. Recently, however, there emerged a novel possibility of obtaining such information by corpus analysis. If the corpus being analyzed is large enough and contains speech material of spontaneous speech ranging in various speaking styles, the results of the corpus analysis would be virtually equal to the direct observation of real speech

behavior. Wells (2003) stated, in this respect, that such use of speech corpora is something impossible at the time of LPD and far from easy today. The first half of this statement is *sine dubio* true. The easiness, or difficulty, however, varies considerably depending on the design of speech corpora that are available for analysis. In the rest of this paper, I will examine if it is possible to extract valuable information of word-form variation from the corpus known as *Corpus of Spontaneous Japanese*.

2. Corpus of Spontaneous Japanese

2.1 Contents of the CSJ

The *Corpus of Spontaneous Japanese*, or CSJ, is a large-scale, richly annotated corpus of spontaneous Japanese constructed by the collaboration of the National Institute for the Japanese Language, the National Institute of Communications and Technology, and the Tokyo Institute of Technology in the years of 1999-2003 (Maekawa, Koiso, Furui, and Isahara, 2001; Maekawa, Kikuchi, and Tsukahara, 2004). The CSJ contains speech signal of about 662 hour that involves more than 7.5 million running words. More than 95% of the 3,302 speech material uttered by 1,417 speakers is comprised of more or less spontaneous monologues, and the resulting 5% is devoted for spontaneous dialogue and read-speech.

2.2 Talk types and impressionistic rating

Talks in CSJ can be classified into five 'talk types'. *Academic presentation speech* (APS hereafter) is live recording of presentations in various academic meetings covering humanities, social sciences, and engineering. Simulated public speaking (SPS) is extemporaneous public speaking by layman speakers in front of small friendly audience on various everyday topics like "the most joyful memory of my life" and "the town I live in." The total recording hours of APS and SPS are about 295 and 332 respectively. In addition to these two main monologue types, there are also recordings of public lectures (PL hereafter) on academic topics done in front of large layman audience. Dialogues of CSJ involve interview of APS and SPS speakers on the content of their talks, task-oriented dialogue, and, free conversation. Lastly, read-speech involves reading of two short passages extracted from non-fiction books, and the reading of the 'orthographic transcription' (see 2.3) of the speakers' APS and/or SPS talks. This is called reproduction speech. Because there are clear differences of speech spontaneity and speaking style among these talk types, we can use the talk types as a useful criterion of sociolinguistic analysis (see 5.1 and 7.1). In addition to the variation in talk types, CSJ has an additional criterion for talk classification. Nearly all APS, SPS, and PL were impressionistically rated in terms of their spontaneity and formality at the time of recording using five-scale rating (see 7.1).

2.3 Annotations

All speech materials of the CSJ are transcribed in two different linguistic formats. One of them is called 'orthographic' transcription because the transcription uses both *Kanji* (Chinese logograph) and *Kana* (Japanese syllabary) characters like in ordinary Japanese text. The orthographic transcription is fully annotated in terms of the part-of-speech (POS), and is useful for various information retrieval and natural language processing purposes. The other format is called 'phonetic' transcription and is transcribed solely by means of *Kana* characters.

The phonetic transcription is prepared to show the details of phonetic and phonological variations in speech, hence essential for the analysis reported below. Moreover, 44 hours of CSJ data, containing a half million words, is annotated minutely using labels of phonological segment and ToBI-based intonation labels (Maekawa, Kikuchi, Igarashi and Venditti, 2002). This part, called the *Core*, is annotated in terms of syntactic dependency structure as well (except for samples of dialogues and reproductions). Needless to say, the analyses of segmental and intonation labels is of great interest for the current study, but this will be the theme of separate paper. In the current analysis, information about word-form variation will be extracted using the transcriptions and the POS information.

2.4 Notes on notation

Short digression on notation and terminology is necessary here. In the rest of this paper, braces and slashes are used to denote lexemes and word-forms respectively. Slashes are also used to denote phoneme(s). For Japanese, /H/ is used to denote a long vowel like /aH/ and /eH/. /N/ and /Q/ are used to denote syllabic (more exactly, moraic) nasal and geminate respectively. Phoneme /c/ is an alveolar voiceless affricate. Strings like /ky/ and /cy/ stand respectively for palatalized versions of /k/ and /c/. English glosses are presented in single quotation marks. When necessary, Japanese terminology is presented in double quotation marks. The term *lexeme* is used to denote word at the abstract level. On the other hand, the term *word-form* means word at the level of lower abstraction. The important trait of word-form is that each conjugational form of a conjugational lexeme is treated as a separate word-form. As long as non-conjugational lexemes are concerned, lexemes and word-forms are identical if the lexeme has no variation. Phonological shapes of word-forms are called variants and presented also in slashes. If a word-form has more than two variants, the word-form is said to be subject to variation.

3. TWO TYPES OF VARIATION

Two different types of word-form variation are recorded in the transcription of CSJ. One of them is the variations recorded in the phonetic transcription by use of the tag (W). For example, if the Japanese lexeme {kokoro} (‘*mind*’) is realized something like /kokoH/ under the joint effects of weakening of intervocalic /r/ and subsequent vowel coalescence, the speech is transcribed as (W kokoH ; kokoro), where the element to the left of semicolon stands for the pronounced form and the element to the right stands for the intended word-form (Note in passing that the original *Kana* text is phonologically romanized in this and all other examples). In the similar vein, (W koNpyuHta; koNpyuHtaH) shows that the last long vowel of /koNpyuHtaH/ (‘*computer*’) is realized as a short vowel. Put very simply, this tag is applied to the cases where standard word-form is “corrupted” by the weakened and/or incorrect articulation. The right-hand element of a (W) tag shows so-to-speak standard word-form of the lexeme. This ‘standard’ form coincides with the DICTIONARY FORM of the CSJ’s POS annotation (see the next section) if the lexeme is not a conjugational word, and, in the case of conjugational lexeme, the conjugated form of the lexeme.

Here, it is important to note that not all variations are tagged 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 non-(W) 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. In the rest of this paper I will refer to these variation types as P-variation and M-variation. The former stands for the variations marked by the tag (W), and the latter stands for the variations that are not marked by the tag. P and M stand respectively for 'phonetic' and 'morphological.'

4. DATA USED IN THIS STUDY

CSJ contains 130,951 running variants that belong to P-variation. On the other hand, the number of variants belonging to the M-variation differs depending on the way M-variation is recognized. This section is devoted for the explanation of the way M-variation was recognized in this study.

Among the rich POS information of CSJ, DICTIONARY FORM, LEMMA, PRON, POS, and CONJUGATIONFORM are used to recognize M-variation. DICTIONARY FORM (DF hereafter) is *Kana* notation of the standard word-form of a given lexeme, which corresponds, in many instances, to the direction word of the lexeme in Japanese dictionaries. The DF of a conjugational lexeme is its ending form ("*syuushike"*). LEMMA is *Kanji* and *Kana* notation of a DF. In principle, homonyms share the same DF but differ in LEMMA. PRON, which was extracted from the phonetic transcription, is a string of *Kana* showing the real pronunciation. Lastly, CONJUGATION FORM (CF hereafter) is necessary to determine the expected standard word-forms of conjugational lexemes.

In the case of non-conjugational lexemes, DF and PRON are expected to be identical if there is not variation. Accordingly, it is possible to extract candidates of M-variation by generating a set of word-forms sharing the same LEMMA, DF, and POS, thereby checking if there are more than two PRON. In the case of conjugational lexemes, the candidate set can be obtained in a similar way. But the set of candidates should be generated for each CF, and DF should be conjugated to relevant CF before they are compared to corresponding PRON.

In addition to the principles described above, special attention should be paid to the following factors in order to detect M-variation correctly. Particles /wa/, /e/, and /o/ are orthographically written as /ha/, /he/, and /wo/ in the present Japanese orthography and are represented as such in the DF. The mismatch between the DF and PRON in these particles was neglected. Long vowels /eH/ and /oH/ with underlying phonemic strings of /ei/ and /ou/ were neglected as long as the lexeme were either Sino-Japanese or Native-Japanese (because in these two word

classes, /ei/ and /ou/ are pronounced as long vowels almost always unless a morpheme boundary is there). Variation due to sequential voicing ("*rendaku*") was also neglected because sequential voicing is a matter of compound-word formation, and the focus of the present study is on simplex word (see also the last paragraph of this section).The variation of classifiers (or numeral suffixes) requires the most intricate treatment. Although Japanese classifiers show considerable variations, most of the variations are predictable if we know the combinations of numerals and classifiers, but there are some cases where free variation is observed. For example, the combination of {haci} ('*eight*') and {hon} (a classifiers used to count stick-shaped objects like trees and pencils) is pronounced as either /haQpoN/ or /hacihon/, but the same numerative does not show any variation when it is combined with numerals like {ici} ('*one*') and {ni} ('*two*'). It was hence necessary to cross-classify the combinations and check them all.

The number of variants belonging to M-variation thus recognized was 302,019, and the total number of variants covering both P- and M-variation was 432,970. This is the total size of non-DF variants. Needless to say, these numbers represent the size of running word-forms. The number of different word-forms is 11,379 including both P- and M-variations.

Three important notices are to be pointed out at the end of this section. First, the recognition of the variation differs considerably depending on the choice of DF. This is especially important when we analyze lexemes of which there is no social agreement about standard DF. As we will see later in section 6, the rate of variation of a lexeme could differ considerably depending on the choice of DF. Second, as suggested earlier, the focus of the current study is on the simplex words rather than compounds. The POS annotation of the CSJ was carried out twice using two units of different word length what we call SUW (short unit word) and LUW (long unit word), and it is the SUW that was examined in this study. The principal reason for this is that it was the SUW that was marked with the tag (W). Note, however, that it does not imply that we analyzed only those SUW that were used as simplex words. We analyze all SUW regardless of whether they were simplex or part of a compound; but the unit of analysis was always SUW. For example, the SUW {kami} ('*paper*') appeared as the last half of compound /ori-gami/ ('*origami*'). In this study, /gami/ as a part of compound was also counted as an instance of {kami}, but it was not recognized as non-DF variant of the lexeme {kami} because this variant, in this particular lexeme, was generated by the obligatory application of sequential voicing. Third, P- and M-variations were analyzed in a lump rather than separately. Lumped analysis was favored in this study to grasp the whole view of word-form variation.

5. OVERALL ANALYSIS

5.1 Correlation with talk types

The ratio of variation (i.e. the ratio of the number of running non-DF variants –either P or M-- to the total number of all running word-forms including the DF) changes systematically when it is classified as a function of the type of talks stated earlier in section 2. Table 1 shows the number of total word-forms, number of non-DF variants covering both P- and M-variations, and the rate (%) of variation (i.e. the latter number divided by the former and multiplied by

100). The rate of variation is the highest in dialogue and lowers gradually down to Reading. The most straightforward explanation on this trend is that it is the reflection of the spontaneity of speech and/or contextually determined speaking style of the talks.

Talk Type	N word-forms	N non-DF variants	% Var
Dialogue	153,591	12,037	7.8
SPS	3,657,277	242,779	6.6
PL	286,611	17,976	6.3
APS	3,344,616	152,094	4.5
Reading	210,429	8,355	4.0

Table 1. Talk types and the occurrence rate of variation

5.2 Word-forms with high frequency of variation

Table 2 lists 20 word-forms that showed the highest frequency of 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 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. Note, in Japanese, both {watasi} and {watakusi} are first-person pronoun, but they were recognized as different lexemes in CSJ. The 19th row of table 2 is concerned only with {watasi}.

This table provides us with the following interesting findings. In the first place, the sum of the frequencies of non-DF variants reaches as many as 325,639 and covers 75% of the total number of non-DF variants in the current data. This suggests that we can examine for most part of the word-form variation by analyzing relatively small number of word-forms. In the second place, most of the lexemes listed in table 2 belong to the basic vocabulary of Japanese and includes both content and function words. In the third place, it is to be noted that the occurrence rate of variation shown in the sixth column of the table is not necessarily high for all items in the table. Some items with low rate of variation (like particle {ni} and pronoun {mono}) are listed in the table simply because their total frequency are stupendously high. Lastly, all items in the table were used by many speakers as shown in the last column. This is the number of speakers who used the word-form at least one time, and regardless of the presence of variation.

The rate of variation shown in the sixth column of table 2 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 3 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 2. 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: good news for those who want to compile corpus-based data for the amelioration of traditional pronunciation dictionaries.

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

Table 2. Twenty word-forms of the highest frequency of non-DF variants

5.3 Word-forms with high rate of non-DF variation

As suggested earlier, word-forms of high frequency of variants are not necessarily those of high rate of variation. Table 4 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 2 and 4. Note, in passing, word-forms whose frequencies were fewer than 10 were removed from the computation for this table. Without this operation, the table would have been occupied by those word-forms whose occurrence frequencies were very low (typically just one) and whose observed word-forms were all different from their DF. Also, some lexemes of high frequency were removed from the table for several reasons. For example, {zu}, an auxiliary verb of negation, was removed because its variants included word-forms of classical Japanese, like /nu/ and /zaru/, in addition to the contemporary forms like /N/.

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/

Table 3. Coverage of non-DF variants by top variants (Same order of row as in table 2).

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

Table 4. Word-forms of the highest occurrence rates of non-DF variants

6. INDIVIDUAL ANALYSIS

In the following sub-sections, each item in table 4 will be examined separately. Although this is not a thorough linguistic analysis, some hitherto unknown aspects of word-form variation in Japanese will be revealed.

6.1 {iu} in adnominal form

Variation of the adnominal form of verb {iu}(‘*say*’) between /iu/ and /yuH/ is very well known among Japanese speakers. In CSJ, /iu/ is adopted as the DF following the custom of Japanese dictionaries. Among the 132,818 total occurrences, the DF /iu/ and its sub-variants (/iH/ and /i/) occurred only 195 times, while /yuH/ and its sub-variants (/yu/, /yuu/, /yuN/, /yuQ/, /yuhu/, /uH/, /yuyu/, /yo/, /yoH/, and /yui/) occurred 132,322 times and covered more than 99%. The top three variants were /yuH/ (119,876 times), /yu/ (8,519), and /yuu/ (3,834). Although variants like /yuH/ are very predominant forms in this word-form, it does not follow that /yuH/ is also predominant in conjugation forms other than the adnominal form. /yuH/ and its sub-variants were also predominant in ending form (“*syuushikei*”, which is phonologically the same as the adnominal form in the present-day Japanese), but in subjunctive (“*kateikei*”) and adverbal (“*renyou*”) forms, its ratio were as low as 5.3% and 0.85% respectively.

6.2 {meHN}

This is a loanword from English ‘*main*’. In Standard Japanese (or Tokyo Japanese), it is known that underlying morpheme-internal /ei/ vowel sequence is realized as a long vowel /eH/. It is also known that some loanwords make exception to this phonological principle. But {meHN} is not recognized usually as an exceptional case. Most Japanese dictionaries adopt /meHN/ as its DF, as was the case in CSJ. However, in our current data, /meHN/ occurred only 3 times as opposed to /meiN/ that occurred 159 times (no other variant occurred).

6.3 {niQpoN}

This is the country name of Japan, and probably the most well-known instance of word-form variation in Japanese. All Japanese dictionaries adopt both /nihoN/ and /niQpon/ as the direction words. CSJ adopted /niQpon/ as the DF for no clear reason. In the current data, /nihoN/ occurred 8,045 times and covered more than 97% of the total of 8,242 occurrences. /niQpoN/, on the other hand, occurred 195 times and covered less than 3% of all occurrences. This is a surprising result for most Japanese speakers. My informal questionnaire survey revealed that most Japanese speakers responded that the rate of /niQpoN/ was slightly less than that of /nihoN/, when they were asked to guess the occurrence rates of these variants. This seems to be a good example of the limit of native speakers’ ability to introspect one’s own speech behavior. This issue will be discussed in more detail in section 7.3.1.

An important note should be addressed with respect to the selection of the DF of this lexeme. As suggested earlier and will be discussed later in 7.3.1, there is no rigid social agreement about the authenticity of /nihoN/ and /niQpoN/. The choice of /niQpoN/ rather than /nihoN/ as the DF was an arbitrary choice, so-to-speak. Had we adopted /nihoN/ as the DF instead of /niQpoN/, the rate of non-DF variation of the lexeme would be as low as about 3%, and the lexeme would not have been involved in tables 2 and 4. Similar arbitrariness of DF selection

could be pointed out with respect to, at least, {iu} and {yoi} in table 4 (see 6.1 and 6.10). Needless to say, the arbitrariness of DF does not mean that the analyses presented in this and other subsections are meaningless, but we have to be aware of the possibility that the ranking shown in table 4 differs considerably depending on the choice of DF as long as lexemes like {niQpoN} are concerned.

6.4 {kaNzuru} in adnominal form

This is a verb that means '*feel*'. The adnominal as well as ending forms of this verb have two main variants, /kaNzuru/ and /kaNziru/. The suffix variation between /zuru/ and /ziru/ is shared by Sino-Japanese verbs whose root is one *Kanji* character long (/kaN/, in this case, means '*sense*'). In CSJ, word-forms ending in /zuru/ were adopted systematically as the DF of these lexemes as is the case in many Japanese dictionaries. But /kaNzuru/ occurred only 8 times whereas /kaNziru/ and its sub-variants occurred 265 times and covered 97% of the total occurrence. See also 6.8 below.

6.5 {simyureHsyoN}

This is also a loanword from English '*simulation*'. The most frequent word-form /syumireHsyoN/, with metathesis, occurred 189 times and covered 83% of the total frequency including the DF (227). The second highest variant was /simireHsyoN/ that occured 15 times. The DF /simyureHsyoN/ (7 times) was the third in rank.

6.6 {enueicikeH}

This item stands for three alphabets "NHK", which is an abbreviation of "*Nihon H0s0 Ky0kai*" ('*Japan Broadcasting Corporation*'), Japan's largest broadcasting company. Most dictionaries, including NHK's pronunciation dictionary, adopt /enueicikeH/ as the DF, but this form occurred only 6 times and was fourth in rank. The three top-most variants included /enuecikeH/ (132 times), /eneHcikeH/ (24), and /enueQcikeH/ (9). These three variants cover 94% of total occurrence including the DF.

6.7 {taiiku}

This is a Sino-Japanese noun meaning '*gymnastics*.' The top three variants were /taiku/ (123 times), /taiQ/ (19), and /taiiku/ (6), and covered 98% of the total occurrence. All instances of /taiQ/ occurred in compound words as the first half of {taiiku-kaN} ('*gymnasium*') and {taiiku-kai} ('*sports association*'). However, it does not imply that /taiQ/ is a conditional variant. /taiiku-kan/, for example, could be pronounced as either /taikukaN/, /taiQkaN/, or /taiikukaN/. Frequencies of these word-forms in CSJ were 36, 11, and 3 times respectively.

6.8 {syoHzuru} in adnominal form

This is another example of suffix variation between /ziru/ and /zuru/. As was the case in /kaNzuru/, word-form ending in /zuru/ (/syoHzuru/) occurred only 7 times whereas /syoHziru/ occurred 108 times and covered 93% of the total frequency.

6.9 {poi}

This suffix derives an adjective out of a root noun. For example, when it is attached to

/huryoH/ (‘*delinquent*’), the resulting /huryoHpoi/ (or /huryoH-Qpoi/) means ‘*hoody*.’ /Qpoi/ occurred 139 times and covered 95% of the total frequency, while the DF {poi} occurred 9 times. All instances of /poi/ occurred immediately after /N/ or /H/, but this variant is not completely conditional, because /Qpoi/ also occurred in this phonological context like /nihoN-poi/ and /nihoN-Qpoi/ (‘*Japanesy*’).

6.10 {yoi} in ending form

It is also well known that adjective {yoi} (‘*good*’) varies between /yoi/ and /iH/. Like most Japanese dictionaries, /yoi/ was adopted as the DF. Among the total occurrence of 4,446 times, the DF /yoi/ occurred 335 times, while /iH/ and its sub-variants (/i/, /iQ/, /ii/, and /iN/) occurred 4,100 times. The top three variants were /iH/ (4,051 times), /yoi/ (335), and /i/ (34). Here again, dominant variants differ depending on conjugation forms. In adnominal form, /iH/ (5,118 times), /i/ (99), and /ii/ (4) covered 88% of the total occurrence (5,950), but in adverbal form that occurred 2,188 times, it was the variants of the type /yoi/ --- e.g., /yoku/ (1,211), /yokaQ/ (953), /yoQ/ (4), etc. --- that were predominant, while variants of the type /iH/ occurred only 6 times. In subjective form, that occurred 44 times, all variants belonged to the type of /yoi/ (e.g. /yokere/, /yokeH/ etc.).

7. SOCIAL FACTORS OF VARIATION

Dependence between the talk types and variation was pointed out in section 5.1. Similar analysis can be applied to individual word-forms if they occur frequently. In this section I will present some examples that suggest the potential of CSJ as the resource of social and stylistic study of variation.

7.1 Effect of talk types

Figure 1 shows the ratio (%) of three main variants of adverb {yahari} (‘*after all*’, see table 2), i.e., /yaQpari/, /yahari/, and /yaQpa/, as a function of talk types. The ratio of /yahari/ is the highest in APS, and lowers monotonically from SPS to Dialogue. On the other hand, ratio of /yaQpari/ is the highest in Dialogue, and lowers monotonically from SPS to APS. The behavior of /yaQpa/ is the same as that of /yaQpari/, but its frequency is much fewer than the others. This figure suggests the interpretation that /yahari/ is the most formal of all, /yaQpari/ is intermediate, and /yaQpa/ is the least formal.

Figure 2 shows the relation between the ratio of {yaQpari} and the impressionistic rating of speech spontaneity (see 2.2). General tendency is that the ratio of {yaQpari} correlates positively with the rated spontaneity of talks. At the same time, it is interesting to see that there is so-called ceiling effect in APS, but not in SPS. This is probably because there is social agreement about the lower limit of formality in APS. Similar ceiling effect was found in lexemes other than {yahari} (See Maekawa, Koiso, Kikuchi, and Yoneyama, 2003, and, Maekawa, 2004)

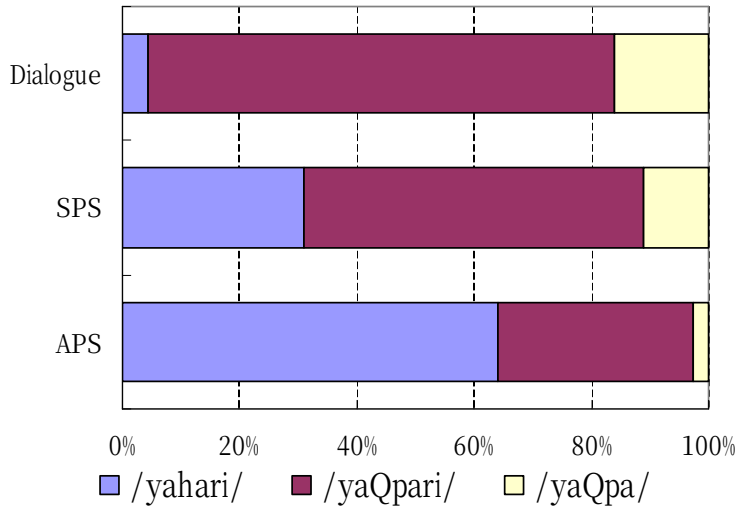

Figure 1. Rates of three main variants of {yahari} as a function of talk types.

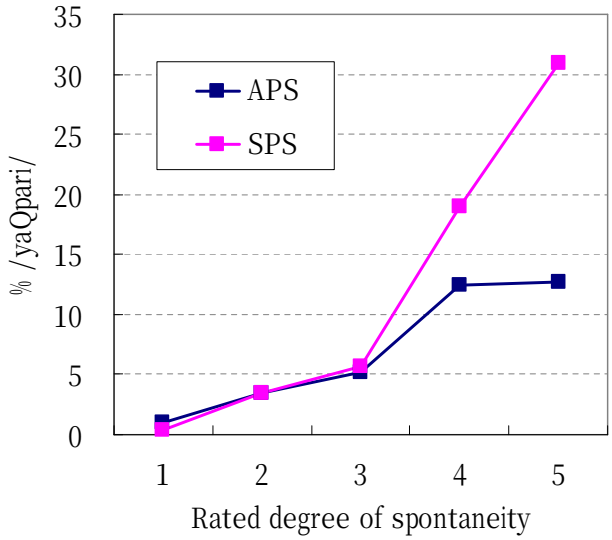

Figure 2. Relation between the rate of /yaQpari/ and the impressionistic rating of speech spontaneity.

7.2 Effect of speaker's age

Figure 3 shows the ratio of {yaQpari} (see tables 2 and 3) as a function of the birth-year of speakers. Age-related systematic difference can hardly be seen in this figure, suggesting that the choice of the three variants of {yahari} is done mostly as a function of speech formality and/or spontaneity. It does not mean, however, that speaker's age is not an important factor of variation. On the contrary, speaker's age could be a very important factor of variation in some lexemes. Figure 4 shows the ratio of /syoHzuru/ in {syoHzuru} (see 6.8). Drastic difference can be seen between the speakers born before and after 1950.

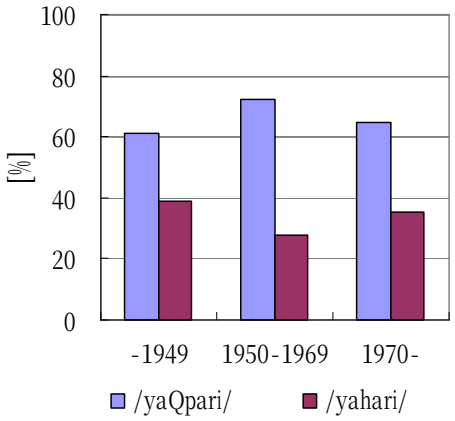

Figure 3. Relation between the rates of two main variants of {yahari} and the birth year of speakers.

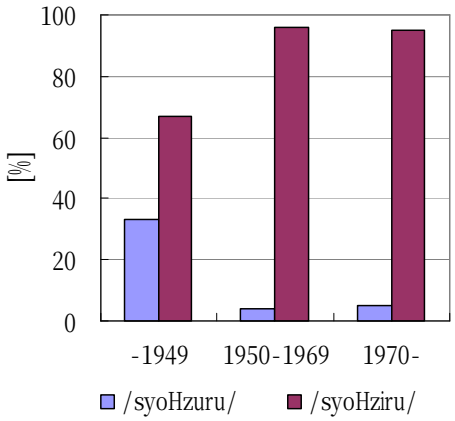

Figure 4. Relation between the rates of two main variants of {syoHzuru} and the birth year of speakers

7.3 Behavior and introspection

As mentioned in introduction, one of the basic motivations for corpus analysis was the belief that the questionnaire was not the best method of knowing people's speech behavior. In this subsection, two examples are to be shown on the gap between people's speech behavior per se and their introspection about the behavior.

7.3.1 Case of {niQpoN}

In section 6.3 we saw that use of /nihoN/ overwhelmed that of /niQpoN/. As suggested there, Japanese speakers are aware that /nihoN/ is more frequent than /niQpoN/, but they do not think that /nihoN/ has overwhelming frequency. According to a questionnaire survey done by NHK Broadcasting Culture Research Institute in 2004, 61% and 37% of subjects responded /nihoN/ and /niQpoN/ respectively when they asked their reading of '*Japan*'. There seems to be several reasons for the overestimation of /niQpoN/. For one thing, announcers of NHK, the most influential broadcasting company in Japan, use systematically /niQpoN/ as the official name of the country. For another, some people believe that there was a governmental decision in favor of /niQpon/ in pre-war Japan, which is an ungrounded legend. At any rate, this is a typical case showing that people's introspection of their own speech behavior could be deadly incorrect sometimes.

7.3.2 Case of potential verb

Variation of potential verbs is a well-known variation of 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 (See Matsuda 1993 for details of this variation). People believe intuitively that the innovative forms like /mi-re-ru/ and /tabe-re-ru/ are now almost predominant in at least young generation's speech. Figure 5 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 6 is the result obtained by analyzing the 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 made report of their norm of writing without knowing it. Use of innovative forms in writings is still exceptional even among the subjects who use innovative forms in their speech rather consistently.

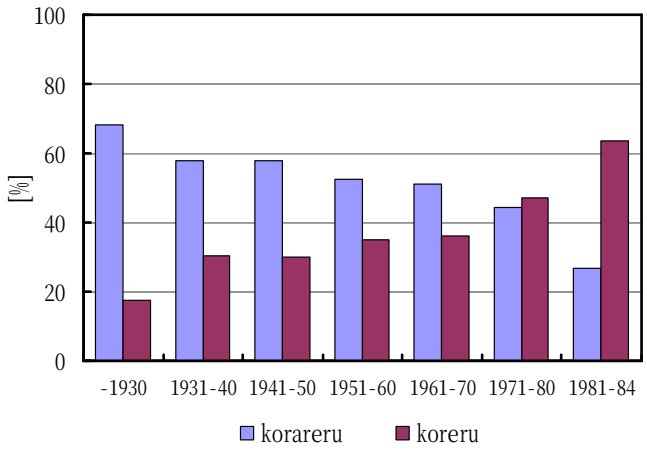

Figure 5. Questionnaire data about the use of potential form of {kuru} as a function of speakers’ birth year.

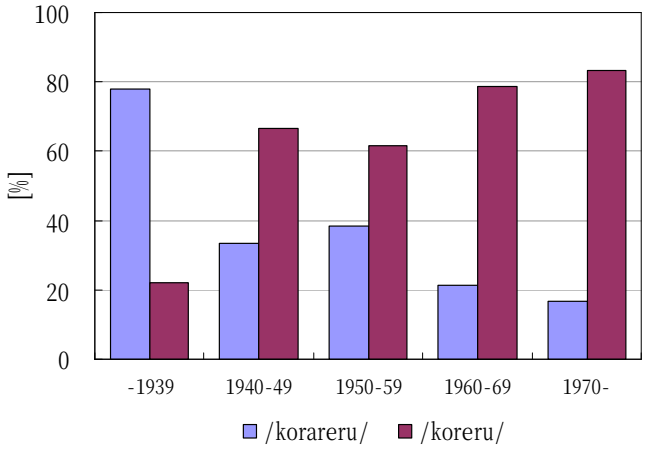

Figure 6. Corpus data about the use of potential form of {kuru} as a function of speakers’ birth year.

8. ESTIMATION OF THE NUMBER OF ANALYZABLE WOD-FORMS

So far, we have seen that CSJ could be an excellent resource for the study of word-form variation. But one important question remains unanswered: How many word-forms could be analyzed successfully by using CSJ? Table 5 simulates the number of analyzable different

word-forms (including the DF) as a function of three parameters: total frequency of the WORD-FORM, ratio of non-DF variants over the total frequency (%), and the number of speakers who uttered the WORD-FORM at least one time. Numbers in the table give the lower bounds of each parameter. So, for example, the last column of the first row is the number of analyzable WORD-FORM under the parameter condition that the WORD-FORM appeared at least 20 times in the corpus, at least 5% of them were non-DF variants, and, the WORD-FORM was uttered by at least 5 speakers. Needless to say, an increase in each parameter decreases the number of analyzable DF.

Freq. WF	% Non-DF Var.	N Speaker	N Analyzable WF
20	5	5	535
20	5	10	325
20	5	50	116
20	10	5	371
20	10	10	239
20	10	50	99
20	20	5	246
20	20	10	171
20	20	50	69
50	5	5	425
50	5	10	294
50	5	50	116
50	10	5	261
50	10	10	208
50	10	50	99
50	20	5	156
50	20	10	140
50	20	50	69
100	5	5	315
100	5	10	247
100	5	50	114
100	10	5	177
100	10	10	161
100	10	50	97
100	20	5	99
100	20	10	96
100	20	50	67

Table 5. Number of analyzable word-forms

It is important to note that among the parameters of this table, the ratio of non-DF variants is not as substantial a parameter as others, because the choice of DF can be an arbitrary selection, and the selection could result in considerable difference in the ratio of non-DF variants as in the cases of {niQpoN, {iu}, and {yoi} (see section 6). Supposing we need at least 100 running word-forms and 50 different speakers to make analysis of social factor like the ones shown in figures 1-4, the maximum number of analyzable word-forms could be 67-114 depending on

the ratio of non-DF variants. Also, supposing that we need at least 50 running word-forms spoken by at least 10 different speakers, 140-294 different word-forms would be analyzable. Incidentally, these numbers are close to the number of items in LPD to which quantitative information were provided. In LPD, by my rough estimation, there are about 200 items with quantitative information, and half of them are shown with figures about social and/or stylistic stratification.

Lastly, one important advantage of corpus-based analysis is to be pointed out. By analyzing corpus we could get information about the absence of variation as well as presence. There are more than ten thousand word-forms in the CSJ that occurred more than 20 times and did not show any word-forms other than the DF. For these, we can say with much confidence that they are not subject to variation. This kind of information can hardly be obtained by questionnaire survey.

9. CONCLUSION

In this paper, I tried to evaluate the usefulness of spontaneous speech corpus as the resource for a pronunciation dictionary. There are four main findings in the current study.

1: Valuable information about the variation of word-forms can be extracted from a large, annotated corpus of spontaneous speech; CSJ in our case.
2: It is also possible to conduct corpus-based analysis about the social and/or situational factors of variation for some word-forms of high frequency.
3: There can be considerable discrepancies between people's speech behavior and their introspection about the behavior gathered by questionnaire.
4: The number of quantitatively analyzable word-forms using CSJ is as many as the items shown quantitatively in LPD.

The conclusion that can be drawn based upon these findings is that it is not only possible but also promising to use CSJ as the resource for the enrichment of pronunciation dictionaries currently available for Japanese. The current study, however, is at a preliminary stage. There are still many investigations to be done. The following ones seem to be the most important.

A: Analysis of the variation of compound lexemes
B: Extensive analysis of all analyzable items encompassing both simplex and compound lexemes
C: Development of a quantitative measure that expresses unequivocally the variability of lexemes (i.e. a measure which is not affected by the arbitrariness of the choice of DF).

As for C:, I recently used entropy as a measure of variability and obtained a result that seemed to be very promising. This study will be reported in a separate paper.

ACKNOWLEDGEMENT

I thank Masaya Yamaguchi, Hanae Koiso, Hideki Ogura, Makiro Tanaka, Toshinobu Mogi, and Tatsuo Miyajima for their comments in various stages of this study. I also thank Caroline Menezes for her comments on English expressions.

REFERENCES

Akita, Y. and T. Kawahara (2005) "Generalized statistical modeling of pronunciation variations using variable-length phone context." *Proceedings of IEEE-ICASSP*, 1, pp.689-692.

Maekawa (2004) "Design, compilation, and some preliminary analyses of the Corpus of Spontaneous Japanese," Yoneyama, K. and K. Maekawa eds., *Spontaneous Speech: Data and Analysis* (Proceedings of the First Session of the 10th International Symposium), National Institute for Japanese Language, Tokyo, pp. 87-107. Available on-line from http://www2.kokken.go.jp/~kikuo/public/KMHP1.html)

Maekawa, K., H. Kikuchi, Y. Igarashi, and J. Venditti (2002). "X-JToBI: An extended J_ToBI for spontaneous speech," *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP2002)*, Denver, pp. 1545-1548.

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 (LKR2004)*, Tokyo Institute of Technology COE-21 Program, pp. 19-24. (Available on-line from http://www2.kokken.go.jp/~kikuo/public/KMHP1.html)

Maekawa, K., H. Koiso, S. Furui, and H. Isahara (2000) "Spontaneous speech corpus of Japanese." *Proceedings of the Second International Conference of Language Resources and Evaluation (LREC2000)*, Athens, pp. 947-952.

Maekawa, K., H. Koiso, H. Kikichi, and K. 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)*, Barcelona, pp. 643-646. (Available on-line from http://www2.kokken.go.jp/~kikuo/public/KMHP1.html)

Matsuda, K. (1993) "Dissecting Analogical Leveling Quantitatively：The Case of the Innovative Potential Suffix in Tokyo Japanese"*Language variation and change*, 5 (1) , pp.1-34.

Wells, J.C. (1990) *Longman Pronunciation Dictionary*. Pearson Education, London.

Wells, J.C. (2003) "Pronunciation research by written questionnaire." *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, Barcelona, pp. 215-218.

Information about the *Corpus of Spontaneous Japanese* is available on-line from http://www2.kokken.go.jp/%7Ecsj/public/ index.html