TOWARD A PRONUNCIATION DICTIONARY OF JAPANESE: ANALYSIS OF CSJ

Kikuo Maekawa

National Institute for Japanese Language and Tokyo Institute of Technology (COE21-LKR)

ABSTRACT

Most existing pronunciation dictionaries do not provide quantitative information about phonological variation of words. Pilot study was conducted to know if it is possible to extract useful quantitative information from the Corpus of Spontaneous Japanese that contains 7.5 million words. 11,379 types of variations containing 432,970 tokens were found in the corpus. It was possible to cover more than 80% of tokens by enumerating the three most frequent variants. For some words, moreover, it was also possible to provide information about the stylistic differentiation among variants based upon quantitative analysis.

1. INTRODUCTION

The aim of this paper is to examine if it is possible to extract from a large corpus of spontaneous speech useful information about the variation of pronunciation, thereby making contribution to the improvement of pronunciation dictionaries. The pronunciation dictionary considered in this paper is mostly for human usage, but the quantitative data reported below should be useful for automatic speech recognition of spontaneous speech, too.

Traditional pronunciation dictionaries provide information about the surface phonological form of the words of a language in qualitative manner. Typical, and basic, format of presentation is a list of variants, which is occasionally accompanied by frequency information in a qualitative manner (e.g. 'Variant A appears more frequently than B'), and/or information about stylistic difference among variants ('variant A is casual while B is formal'), and/or information about age-gradation ('variant A is used mostly by young speakers').

Recently, some pronunciation dictionaries of English, try to provide quantitative information about the variation. *Longman Pronunciation Dictionary* [1] shows, for an example, that in American English, the last syllable of "attitude" is pronounced as /tu:d/, rather than /tju:d/, in more than 85% of cases.

Although this is an obvious advance in pronunciation dictionary, there seems to be room for further advancement. One important limitation of the *Longman* dictionary consists in that the quantitative data was obtained by means of questionnaire survey rather than the direct observation of speech. Although questionnaire is an effective way of gathering coherent data on many words simultaneously, it depends heavily upon the speakers' ability to introspect one's own speech behavior. As will be suggested below, it is not always easy to arrive at correct observation of one's own speech behavior by introspection.

An effective way of observing phonological variation of lexical entries in the real speech behavior consists in the analysis of large-scale corpus of spontaneous speech. In what follows, I present the result of a pilot survey about the phonological variations at the lexical level observed in the *Corpus of Spontaneous Japanese*.

2. CORPUS OF SPONTANEOUS JAPANESE

Corpus of Spontaneous Japanese, or CSJ, contains speech signal of 662 hours involving about 7.5 million words [2-4]. It involves 3,302 speech files spoken by 1,418 different speakers. 95% of the speech material is spontaneous monologue. Resulting 5% is devoted for spontaneous dialogue and read speech.

Speech material of CSJ can be classified into five talk types. *Academic presentation speech* (APS) is live recordings of academic presentations done in the meetings of various academic societies covering both humanity and engineering studies. *Simulated public speaking* (SPS) is laymen's public speaking on everyday topics done in front of small audience without written-to-be-spoken manuscript. APS and SPS cover 95% of CSJ. In addition to these, there are also three talk types: spontaneous *Dialogue* (D), *Reading text* (R), and *Miscellaneous* (M). The last type includes academic presentations done by specialists in front of large layman audience. The crucial difference between APS and M consists in the size and character of the audience rather than the contents of talks. The types D, R, and M were recorded in the corpus mainly for the comparison with the APS and SPS materials.

All materials are transcribed and morphologically annotated. Segmental and prosodic annotations are provided for a subset corpus, called the *Core*, that contains about a half million words (corresponding to about 44 hours). The transcriptions of CSJ, as well as segmental and prosodic annotations, are excellent resource for the quantitative study of phonological variation. In the rest of this paper, I will present the results of the analysis of transcription text. Analysis of segmental and prosodic annotation will be the theme of separate paper.

3. TWO TYPES OF VARIATIONS

In the transcription of CSJ, a special tag '(W)' was used to represent phonological variations. For example, /kokoro/ ('*mind*') is occasionally pronounced like /kokoH/ (where /oH/ stands for a long vowel) due to the considerable weakening of /r/ franked by two /o/s. Cases like this were annotated as "(W kokoH; kokoro)", where the strings to the left and right of semicolon represent the pronounced and canonical forms respectively. 'Canonical' forms correspond usually to the entry words in dictionaries. In the same vein, (W koNpyuHta; koNpyuHtaH) records the shortening of lexical long vowels in the last syllable of a loan word meaning '*computer*'.

It is not the case, however, that the (W) tags mark all variations that are relevant to this study. There are many variations to which the tag was not applied. Examples are two variants of the country name of 'Japan', /nihon/ and /niQpon/, two variants of the verb 'say', /iu/ and /yuH/, and two variants of first-person pronoun 'I', /watasi/ and /atasi/. None of these variants were marked by (W), because it is impossible to determine which of the two is the canonical form. Roughly speaking, (W) is applied to the variations that are either sporadic or caused by careless articulation, while the tag is not applied to the variations that most speakers are aware of. Accordingly, many of the non-W variants including the ones cited above are adopted as entry words in ordinary Japanese dictionaries.

4. DATA

The CSJ contains 130,951 cases of variations marked by (W), which are referred to as 'W variant' hereafter. As for the variations that are not marked by (W), which will be referred to as 'non-W variants,' the occurrence frequency differs depending on the way we define variation. In this study, variations are defined as follows.

Among the morphological information of CSJ are LEMMA, DICFORM, and PRON. For any given word, LEMMA is a string of Kanji and Kana that shows canonical orthographic form of the word. DICFORM (standing for 'dictionary form') also shows the ordinary orthographic form but as a string of Kana. PRON, on the other hand, shows the pronounced form of the word. Put very simply, if a class of words defined by the same set of LEMMA, DICFORM, and POS (part-of-speech) specifications has more than two PRON forms without the (W), the word has variation, and the number of different PRON is the number of non-W variants.

It is important to note that the POS specification used in the survey included information about the conjugation forms (like attributive form, ending form, and so on), because without this specification, every different conjugation form are recognized as variant. Under this definition, CSJ contains 302,019 non-W variations, and the total frequency of variations becomes 432,970. Needless to say, these frequencies are on token-basis. The type-basis frequency is 11,379.

5. ANALYSES

5.1. Overall tendencies

Before examining individual words, two interesting overall tendencies are to be mentioned. First of all, the number of types having variation (i.e., 11,379) corresponds to about 20% of all types in the CSJ (i.e., 56,618 when counted in the same manner as adopted for the counting of non-W variants). This suggests strongly that, unlike the intuitive belief of some linguists, pronunciation of spontaneous Japanese is not very stable.

The second tendency to be noted is the dependence of variation rate upon the talk types. As shown in table 1, occurrence rate of variants (both W and non-W) is the highest in D, and lowers gradually in the order of D>SPS>M>APS>R. This order is most likely to be that of the degree of speakers' consciousness about the way they speak, i.e., the spontaneity of talks with D and R being the highest and lowest.

Table 1. Rate of variation as a function of talk types.

Talk type	N word	N variation	% Var
D	153,591	12,037	7.8
SPS	3,657,277	242,779	6.6
M	286,611	17,976	6.3
APS	3,344,616	152,094	4.5
R	210,429	8,355	4.0

5.2. Words with high frequency of variation

Table 2 shows 20 words whose frequency of variation are the highest. Variations associated to these words cover 75% of the total number of observed variations (325,636 out of the total of 432,970). It can be seen from the table that both content and function words are in the table. It is also to be noted that the probability of variation (the fifth column; the total number of variation divided by the total number of the word in CSJ) is not necessarily high for all words. And, variations are observed in many speakers.

The last column of the table shows the number of speakers who showed at least one instance of variation.

Note, at this point, that ‘N Var’ in the fourth column of the table is the total number of observed variations and includes many different variation types. The second column of table 3, on the other hand, shows the number of variation types observed for the words in table 2. The third column of table 3 shows the percentage of the total frequency of variation covered by the frequency of the most frequent variant. Likewise, the fourth column shows the percentage covered by the sum of the frequencies of three variants whose frequencies are the highest. The last column shows the phonological forms of the most frequent variants.

Table 3 shows that for 10 words out of 20, the sum of the three most frequent variants cover more than 99% of the observed variations; there are only two words (/sore/ and /desu/) whose coverage by the three most frequent variants are less than 80%. This table suggests strongly that it is not necessary for a pronunciation dictionary to show a long list of variants even in the case of words with the highest frequency of variation. It is possible to compile a pronunciation dictionary that covers about 80% of the variations by making list of three most frequent variants.

Table 2. Words that showed the most frequent variations. ‘N Spk’ is the number of different speakers who used variants. See text for other columns.

LEMMA	DICFORM	POS	N Var	Prob. Var	N Spk
言う (連体)	iu	Verb (attrib)	120,854	0.91	1411
の (準体助詞)	no	Nominalization	79,829	0.52	1326
けれど	keredo	Particle	26,534	0.56	1092
何	nani	Pronoun	17,140	0.74	1054
言う (終止)	iu	Verb (ending)	7,991	0.87	1031
って	Qte	Particle	7,834	0.16	956
日本	niQpon	Noun	7,801	0.95	849
くらい	kurai	Particle	7,758	0.87	951
に (格助詞)	ni	Case particle	7,568	0.04	1097
矢張り	yahari	Adverb	7,022	0.60	706
其れ	sore	Pronoun	6,016	0.14	767
良い (連体)	yoi	Adjective (attrib.)	5,177	0.87	934
良い (終止)	yoi	Adjective (ending)	4,026	0.91	866
もう	moH	Adverb	3,669	0.20	674
です	desu	Aux. Verb	3,431	0.02	624
で	de	Conj.	3,290	0.06	756
皆	mina	Noun	2,634	0.61	593
物	mono	Noun	2,373	0.08	593
私	watasi	Pronoun	2,367	0.15	395
そう	soH	Adverb	2,327	0.08	619

Table 3. Number of variation types and the percentages of the number of variations covered by the frequency of the highest frequency variation types. The last column shows the word form of the most frequent variant (MFV).

DICFORM	N Var Type	%Coverage (1)	%Coverage (1-3)	MFV
iu	96	92.4	99.6	yuH
no	80	99.4	99.8	N
keredo	124	82.6	98.0	kedo
nani	65	97.2	99.2	naN
iu	20	94.0	99.3	yuH
Qte	97	50.5	85.9	te
niQpon	13	99.3	99.4	nihoN
kurai	18	99.7	99.8	gurai
ni	78	87.0	95.5	N
yahari	136	73.9	90.8	yaQpari
sore	260	38.2	61.6	soe
yoi	12	97.9	99.8	iH
yoi	20	98.7	99.8	iH
moH	36	96.3	97.8	mo
desu	151	38.1	76.7	es
de	74	51.0	89.8	Nde
mina	11	99.5	99.7	miNna
mono	57	88.6	94.9	moN
watasi	76	73.2	92.8	atasi
soH	112	83.1	89.3	so

Table 4. Words with high probability of variation. ‘R’ is the rank in table 2. ‘%Var’ is the ratio of observed variations to the total occurrence of the word in CSJ. See text for other columns.

R	LEMMA	DICTFORM	POS	N Var	Prob. Var
30	ず	zu	Aux. Verb	1,279	0.979
121	メーン	meHN	Noun	157	0.975
97	シミュレーシ	simyureHsyoN	Noun	219	0.965
98	つう	tuH	Aux. Verb (ending)	217	0.964
115	NHK	enueicikeH	Noun	17	0.962
130	体育	taiiku	Noun	145	0.960
7	日本	niQpoN	Noun	7,801	0.947
144	ぽい	poi	Suffix	133	0.917
166	生ずる	syouzuru	Verb (attirb)	106	0.914
1	言う	iu	Verb (attrib.)	120,854	0.910

5.3. Words with high probability of variation

As suggested earlier, words with high frequency of variation are not necessarily the words with high probability of variation. Table 4 shows 10 words with the highest probability of variation (see 5.2). Tables 2 and 4 share only two words. The leftmost column 'R' of table 4 is the ranking of the word when it is put in table 2.

Here, it is important to note that the words whose frequency in the whole CSJ was less than 100, or whose frequency of variation was less than 5 were excluded from table 4. Also, variations that were observed only in one speaker were excluded. Without these restrictions, the top of the list would be occupied by words that occurred only once and whose PRON form was different from the DICFORM.

In the following subsections, each item in table 4 will be observed in detail.

5.3.1. /zu/

Three most frequent variants of the word were /N/ (739 times), /zaru/ (163), and /nu/ (112). These variants cover 99.7% of the total number of variations. Of these, all /N/ are not necessarily free variation, because, for example, /zu/ becomes regularly /N/ after the politeness adverb /masu/ like /tabemaseN/ ('*I don't eat.*'). Excluding all cases following /masu/, there remain 92 cases of /N/, which can be regarded as a true case of variation. On the other hand, /nu/ and /zaru/ are archaic counterpart of present-day /zu/, and are used typically in more or less fixed, or fossilized, expressions like /sezaruo enai/ ('*have no choice but to do something*') or /nitemonicukanu/ ('*far from being similar*').

5.3.2. /meHN/

This is a loan word of English '*main*'. Many Japanese dictionaries adopt /meHN/ as its primary entry word form, because in standard Japanese sequence of two vowels /ei/ is usually pronounced as a long vowel /eH/. But when they are in loan words, /ei/ are often pronounced as they are. In the case of /meHN/, /mein/ appeared 157 times, while the canonical /meHN/ appeared only 3 times. This tendency is observed consistently in most of /ei/ sequences in loan words (but see 5.3.5 below for an exceptional case).

5.3.3. /simyureHsyoN/

This is also a loan word from English '*simulation*'. /syumireHsyoN/ and /simireHsyoN/ occurred 189 and 15 times and covers 93% of all variants. Canonical /simyureHsyoN/ appeared only 8 times in the CSJ.

5.3.4. /tuH/

This word is hardly found in ordinary Japanese dictionary, because it is itself a colloquial variant of word sequence /to iu/ ('*so called*'). In this case, the top 3 variants, namely /QcyuH/ (50 times), /QcuH/ (36), and /cuH/ (28), cover only 53% of total variation. It is necessary to sum up to the eighth variants to cover more than 90%. The additional variants are /cu/ (25), /Qcu/ (18), /teH/ (15), /Qcyu/ (15), and /cyuH/ (10).

5.3.5. /enueicikeH/

This word is made of three alphabets, '*NHK*', and is abbreviation for '*Nihon Hoosoo Kyookai*' ('*Japan Broadcasting Society*'). Top 3 variants, /enuecikeH/ (129), /eneHcikeH/ (24), and /enueQcikeH/ (9) cover 94% of all variation. It is interesting to see that the /ei/ sequence involved in alphabets 'K' (/kei/) is realized not as it is but a long vowel. This is an interesting exception to the tendency described in 5.3.2 above.

5.3.6. /taiiku/

This is a Sino-Japanese word meaning '*gymnastics*.' The most frequent variant is /taiku/ (115 times), in which the long vowel or the sequence of two identical /i/ vowels in the second syllable is shortened. The variant of the second highest frequency /taiQ/ (18) all appeared in compound noun of /taiikukaN/ ('*gymnasium*') and /taiikukai/ ('*gymnastic society*'). The canonical /taiiku/ appeared only 6 times in the whole CSJ.

5.3.7. /niQpoN/

Variation of the name of country '*Japan*' is perhaps the most well known variation in Japanese. More than 99% of the total variation, and 94% of all occurrence in the CSJ is covered by a single variant /nihoN/. It is surprising that /niQpoN/ appeared only 193 times in the whole CSJ.

This fact demonstrates convincingly the necessity of quantitative description in pronunciation dictionaries. As long as I know, all Japanese dictionaries show both variants, and most adopted /nihoN/ as the dictionary form, but none note that /nihoN/ is used by far more frequently than /niQpon/. This is true in the most famous pronunciation dictionaries of Japanese like *NHK*'s and *Sanseido*'s [5, 6].

The lack of quantitative information about /nihoN/ and /niQpoN/ in the existing dictionaries suggests the difficulty of getting quantitative information of language behavior by introspection. An informal survey that I conducted using about 20 persons (mostly my colleagues) as subjects suggested that it was very difficult to estimate the occurrence rate of these variants. Most subjects said that /nihoN/ was more frequent than /niQpoN/, but when I asked them to guess the percentage of /nihoN/ their answers ranged between 60 to 70% in most cases. Perhaps, it is noteworthy that the only subjects who correctly guessed that /nihoN/ was used more than 90% of cases were not specialists of Japanese but a physicist and a government officer.

5.3.8. /poi/

This suffix meaning "like" was pronounced mostly as /Qpoi/ (129 times). Among the 4 cases of /poi/, one was a meta-linguistic usage in an APS about Japanese grammar. The resulting three cases appeared after the so-called special morae, i.e., moraic nasal /N/ (two times), and long vowel /H/. This phonological context, however, is not deterministic, because, /Qpoi/ was also used in the same context.

5.3.9. /syoHzuru/

This is a verb meaning '*yield*' or '*occur*', and is also a well known case of variation. The verb was pronounced as /syoHziru/ in more than 90% of cases.

Morphologically, the verb is derived from a Sino-Japanese nominal stem /syoH/ by attaching a native suffix /zuru/. In Japanese, there are many verbs of this type like /siNzuru/ ('*believe*') and /kanzuru/ ('*feel*'). These verbs share the suffix variation between /zuru/ and /ziru/.

5.3.10. /iu/ in adnominal form

/yuH/ (111,705 times) covers more than 92% of variants, and more than 90% of all tokens in the CSJ. Also, /yu/, the variant of second highest frequency (8,524), covers 7.1 and 6.4% of all variations and all CSJ tokens.

It is important to note, however, that /yuH/ and /yu/ are not as predominant as in adnominal form when other conjugation forms are concerned.

In adverbal form (*renyoukei*), for example, 177 variations were observed out of the total of 20,719 tokens in the CSJ. The most frequent variant was /i/ (94 times). /yu/ (15) and /yuH/ (11) were the second and third in frequency ranking. On the contrary, /iH/ and /iQ/ (phonologically conditioned variants of /iu/ in adverbal form) appeared 16,957 times and shared 81.8% of all tokens.

These facts suggest strongly that conjugation form is an important factor of variation in conjugation words like verbs and adjectives.

5.4. Factors of variation

5.4.1. Linguistic factors

It is well known that linguistic variations are influenced both by linguistic and social, or extra-linguistic, factors [7].

Full-fledged analysis of linguistic factors is beyond the scope of this paper, but some linguistic factors were mentioned in the preceding subsections; namely, preceding word, fixed phrase (fossilization), word origin (i.e., loan, Sino-Japanese, and native), abbreviation using alphabets, word-compounding, and, phonological context. Here, it is important to note that most of the phonological difference observed between the dictionary form and the MFV column of table 3 are related to the insertion/ deletion of so-called special morae (i.e, /N/, /Q/, and /H/), or deletion of /r/ consonant. It might be the case that most part of the variations observed in spontaneous speech can be explained by the combination of small number of phonological rules. The quantitative analysis of this hypothesis will be presented in a separate paper.

5.4.2. Social factors

We have already seen in section 5.1 that talk type has systematic influence on the overall ratio of variation. Needless to say, this kind of analysis can be applied to individual words. Figure 1 shows the percentages of /yahari/, /yaQpari/, and /yaQpasi/, the three main variants of canonical /yahari/ (which means '*also*' or '*either*', see tables 2 and 3), as a function of talk types. It is clearly seen that the percentages of canonical, and putatively the most formal, /yahari/ diminishes distinctively from APS to SPS, and, from SPS to Dialogue. On the contrary, colloquial /yaQpari/, and more colloquial /yaQpasi/ increases as the formality of speech lowers from APS to SPS and Dialogue.

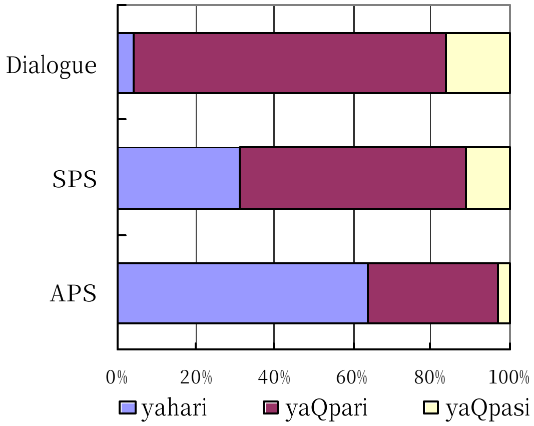

Figure 1. Percentage usage of the three variants of /vahari/ as a function of talk

In addition to the analysis in terms of talk type, it is also possible to analyze data as a function of the information about individual talks and speakers. Figure 2 shows the relationship between the percentage usage of /yaQpari/ and the rated degree of spontaneity of talks. By rated degree of spontaneity is meant an overall impressionistic evaluation of individual talks given by naïve raters. CSJ provides various impressionistic evaluation data including 'degree of spontaneity', 'formality of speaking style', 'subjective speaking rate', and so forth.

In the abscissa of figure 2, higher number stands for higher spontaneity, i.e., the talk in question is spoken without reference to prepared text and impromptu.

The percentage shared by /yaQpari/ increases both in APS and SPS as rated degree of spontaneity increases. But it is interesting to see that the two line diverse at the right end. This is presumably a ceiling effect of speaking style. Although there is a correlation between the speaking style and spontaneity, the correlation is not perfect because one cannot lower the formality of talk unlimitedly in the situation of academic presentation even if the talk is ultimately spontaneous. It seems that this constraint takes strong effect when a young speaker gives a talk is front of large audience who are mostly older than the speaker: the most frequent situation of APS talks.

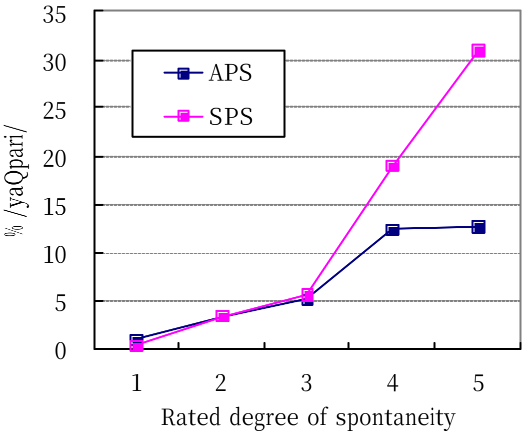

Figure 2. Ratio of variant /yaQpari/ to the total of all tokens in the CSJ as a function of the rated degree of spontaneity.

6. PROSPECTS

The results of pilot survey show that it is not only possible but also promising to use spontaneous speech corpus as the source of quantitative information for the improvement of pronunciation dictionary. Figure 3 shows an example of quantitative description of variation in an improved pronunciation dictionary. In the first line of each lexical entry, the heading is followed by canonical phonological form(s), if there is any. Frequent variants are shown below the heading with their relative frequency in the whole CSJ and note on stylistic differentiation, when necessary. Two entries have NB about minor phonetic variations observed in the CSJ. This will be useful for those who are learning Japanese as a foreign language.

The information that can be extracted from CSJ is not limited to the ones in figure 3. It is possible to extract data about the variation of lexical pitch accent from the *Core* of CSJ.

Lastly, astute readers may want to know the number of words to which this kind of information can be provided using CSJ. Under the criterion of data selection used in section 5.3, we obtain 1,528 different words. If we adopt another criterion that involves words whose CSJ frequency is higher than 49 and whose probability of variation is equal to or higher than 0.1, we obtain 3,023 words.

Quantitative information of variation improves greatly the descriptive power and usefulness of existing pronunciation dictionaries. Results of the ongoing survey about the variations of lexical pitch accent should provide further enrichment of the dictionaries.

NHK (noun) *enueichikee*
enuechikee (72%)
eneechikee (13%)
enuecchikee (5%)
enueichikee (3%)

日本 (noun) *nihon, nippon*
nihon (94%)
nippon (4%)
N.B. 1) No stylistic difference between *nihon* and *nippon*. 2) *nihon* may sound like *nion*.

矢張り (adv.) *yaha'ri*
yappari (46%)
yahari (31%)
yappa (8%) Casual
yappashi (1%) Casual
N.B. *yappari* may sound like *yappai*.

Figure 3. Quantitative information in a pronunciation dictionary.

ACKNOWLEDGEMENT

I thank Makiro Tanaka of NIJL who let me know about the distribution of three variants of /yahari/ shown in figure 1.

REFERENCES

[1] Wells, J. C. *Longman Pronunciation Dictionary (New Edition)*. Pearson Education, Harlow, 2000.

[2] Maekawa, K. "Corpus of Spontaneous Japanese: Its Design and Evaluation", *Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition* (SSPR2003), Tokyo, pp.7-12, 2003.

[3] Maekawa, K., H. Kikuchi, and W. Tsukahara, "Corpus of Spontaneous Japanese: Design, Annotation and XML Representation", *Proceedings of the International Symposium on Large-scale Knowledge Resources* (LKR2004), Tokyo Institute of Technology, Tokyo, pp.19-24 ,2004.

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

[5] NHK Housou Bunka Kenkyuujo. *NHK Nihongo Hatsuon Akusento Jiten*. NHK Shuppan, Tokyo, 1998.

[6] Akinaga, K. *Shinmeikai Nihongo Akusento Jiten*. Sanseido, Tokyo, 2001.

[7] Labov, W. *Principles of Linguistic Change*. (2 Vols.) Blackwell, Oxford, 1994-2001.