USE OF A LARGE-SCALE SPONTANEOUS SPEECH CORPUS IN THE STUDY OF LINGUISTIC VARIATION

Kikuo Maekawa, Hanae Koiso, Hideaki Kikuchi and Kiyoko Yoneyama
{kikuo, koiso, kikuchi, yoneyama}@kokken.go.jp

Dept. Language Research, The National Institute for Japanese Language, Tokyo

ABSTRACT

Corpus of Spontaneous Japanese, or CSJ, is a large-scale database of spontaneous Japanese. It contains speech signal and transcription of about 7 million words along with various annotations like POS and phonetic labels. After describing its design issues, the potential of the CSJ as a resource for linguistic variation study was evaluated.

1. INTRODUCTION

Study of linguistic variation is one of the central issues in sociolinguistics. There is a wide agreement among the researchers that the study should be based upon observation of naturalistic data rather than the data gathered by questionnaire. In this respect, speech corpus containing large amount of spontaneous speech can be an excellent resource for the study of language variation. There is, however, another kind of consensus among sociolinguists that the observation should encompass different speaking styles, because stylistic variability can give us the possibility of observing linguistic change in progress [1]. Accordingly, spontaneous speech corpus has the potential of being an excellent resource of variation study on condition that it covers materials of different speaking styles, and/or, there is a way to evaluate he difference of speaking style in an objective way.

Since 1999, the National Institute for Japanese Language --in collaboration with the Communications Research Laboratory and the Tokyo Institute of Technology-- has been compiling a large-scale corpus of spontaneous Japanese called the *Corpus of Spontaneous Japanese* (CSJ, hereafter). Although the primary aim of this corpus is speech recognition study, the corpus is also designed for the study of language variation [2,3].

2. CORPUS OFSPONTANEOUS JAPANESE

CSJ contains about 640 hours of spontaneous speech corresponding to about 7 million words. The main body of the corpus consists of two different types of spontaneous monologue, namely, Academic Presentation Speech (APS) and Simulated Public Speech (SPS).

APS is live recording of academic presentations done in the meetings of 9 different academic societies encompassing the fields of engineering, language science, and social science. On the other hand, SPS is the public speech given by layman speakers on every-day topics like 'my most delightful memory' or 'the town I live in' in front of small number of friendly audience. Needless to say, our expectation was to capture relatively high and low speaking styles in APS and SPS respectively. To achieve this objective, the recording stuff of the SPS tried to relax the speaker as much as possible by spending some time chatting with the speaker prior to the recording.

In addition to the extrinsic difference between the two speech types, speaking style and spontaneity of each individual speech were evaluated impressionistically using a five-category scale (See section 3).

Currently, we have 1007 APS (of about 299 hours) spoken by 805 different speakers and 1715 SPS (of about 330 hours) spoken by 590 different speakers. All speech materials are divided into transcription units at the locations of longer-than 200ms pauses and transcribed using a special Kana-Kanji orthography devised for CSJ. The transcribed text is annotated with respect to its verbal and non-verbal characteristics including filled-pauses, word-fragment, reduced articulation, laughter, and so forth. At the same time, the transcribed text was POS (part-of-speech) analyzed.

In addition to these annotations, segmental and intonational annotations are provided for the material of about 44 hours (about 500k words) constituting the true subset of the whole corpus. We call this subset the Core of CSJ. The annotation of the Core is currently underway.

In the rest of this paper, we will present the results of preliminary analyses of the linguistic variation recorded in the corpus.

3. SEGMENTAL & MORPHOLOGICAL VARIATIONS

3.1 Variables and Factors

The linguistic variables analyzed in this section include 1) Devoicing of close vowels (abbreviated as DV hereafter), 2) Shortening of lexical long vowels (SLV), 3) Coalescence of /de/+/wa/ into /zya/ (ZYA1 and 2), and 4) Moraic nasalization of particle /no/ (NO1 and 2). Concise linguistic descriptions of these variables are given below.

ISBN 1-876346-48-5

DV: Close vowels, /i/ and /u/, tend to be devoiced when they are preceded and followed both by voiceless consonants [4]. DV does not affect word meaning.

SLV: Lexically specified long vowels are shortened occasionally. This variation can affect word meaning in the minimal pairs like /oHbasan/ (surname) versus /obasaN/ ('aunt'), where /H/ stands for a long vowel [5].

ZYA: Word sequence of /de/+/wa/ sometimes coalesces into /zya/. Since the POS of /de/ can be either case particle or auxiliary verb, they are distinguished as ZYA1 (particle) and ZYA2 (auxiliary verb)[6].

NO: Particle /no/ can be realized as a moraic nasal /N/. Depending on the subclass of the /no/ particle, we make distinction between NO1 (case particle) and NO2 (nominalization particle) [7].

In addition to these segmental and morphological variations, we will analyze prosodic variation in section 4.

As for the factors of linguistic variation we examine the following ones in this paper.

Type: Either APS or SPS.
Styl: Impressionistically rated speaking style.
Spnt: Impressionistically rated spontaneity.
SRate: Speaking rate normalized within a subject.
Laugh: Occurrence of laughter in a transcription unit.
Sex: Sex of the speaker

Type, Style, and Spnt are the variables incorporated at the time of corpus design. SRate, Laugh, and Sex, on the other hand, turned out to be effective factors during the course of the pilot analysis of the CSJ [4-7].

SRate was computed for each transcription unit as the number of morae spoken per second, and then classified into four categories so that each category involve 25% of the number of transcription units. SRate1 and SRate4 are the slowest and fastest respectively.

Styl ranges from 1 (very casual) to 5 (very formal), and, Spnt ranges from 1 (very spontaneous) to 5 (very prepared).

3.2 Analysis of Main Effects

We applied one-way ANOVA for the six variables to see if there was significant relationship between the variables and factors mentioned above. The results are summarized in Table 1 at the end of this paper. All factors turned out to be significant for at least more than two variables, and, all variables were affected by at least more than five factors. Note, at this point, that the levels of significance are unusually high in Table 1 (e.g. **** stands for P<.0001). This is necessary as well as desirable because the umbers of observation (N) are unusually large in most of the variables.

Needless to say, significance of ANOVA does not imply linguistically meaningful relationship between the variables and factors. But in our data, most of the relationships seem to be linguistically meaningful.

For example, Figure 1 shows the relationship between the variable DV and factor SRate. For both vowels, devoicing rate increased monotonically as a function of speaking rate: a natural relationship for a purely phonetic variation like DV.

Also, Figure 2 shows the relationship between the two variables of ZYA and the factor Styl. It is interesting to see that despite the large difference of the probability of coalescence due to POS difference, both two variables correlated well with the factor Styl.

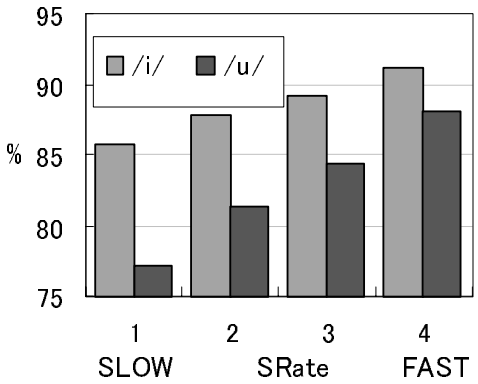

Figure 1. Effect of speaker-normalized speaking rate (abscissa) on the rate of close vowel devoicing (ordinate).

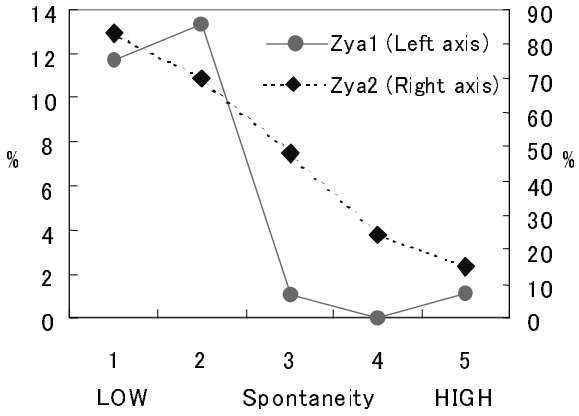

Figure 2. Correlation of the rate of word coalescence (/de/+/wa/>/zya/, ordinate) and impressionistically rated spontaneity.

3.3 Analysis of Interaction

We did not pay attention for the possibility of interaction among factors in Table 1 for the sake of simplicity, but there were many complex interactions in the data. Here, we show two interesting cases as example.

3-way ANOVA of SLV using Sex, Type, and Laugh as the factors revealed significant interactions between Type*Sex and Type*Laugh all at p< .01 level. More interestingly, none of the three main effects was significant. As shown in Figure 3, presence of laughter enhances SLV in SPS but not in APS, and, females are more sensitive to the type of speech than males. These interactions suggest that speakers' relaxedness (which is reflected indirectly in the frequency of laughter) is a hidden, but fundamental, factor of SLV.

The second example is about ZYA. Figure 4 shows the interaction between Laugh and Type with respect to ZYA1 and 2. Interaction in ZYA1 is similar to that observed in SLV (Figure 3A), but the interaction in ZYA2 is different. As long as ZYA1 is concerned, laughter enhanced coalescence in SPS but not in APS, while laughter strongly enhanced coalescence in APS but not so much in SPS when ZYA2 is concerned.

This subtle difference seems to be related to the difference of the relationship between Styl and two sort of ZYA shown in Figure 2. In Figure 2, the curve for ZYA1 (particle) shows step-like abrupt change between Styl 2 and 3, but the curve of ZYA2 is gradual. In addition, there is a large difference of the occurrence probability between ZYA1 and 2 as indicated in the two ordinates of the figure.

These facts suggest the possibility that the occurrence of ZYA1, which is much less probable than ZYA2, is a strong indicator of the lowering of speaking style. If this is true, it is natural that in APS, where style is expected to be not too low, coalescence of ZYA1 was nearly inhibited, but ZYA2 was not, because coalescence of ZYA2 does not run the risk of indicating style lowering.

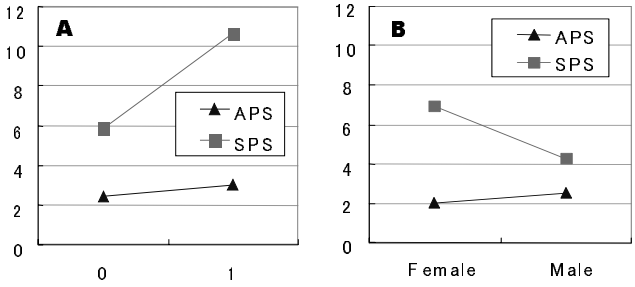

Figure 3. Interaction between Laugh and Type (panel A), and, Sex and Type (B) in SLV. '0' and '1' in panel A denote absence and presence of laughter. The ordinate shows the rate of long vowel shortening [%].

4. PROSODIC VARIATION

In addition to the segmental and morphological variations examined in the previous section, it is possible to examine variations of prosodic phenomena. The extended J_ToBI (X-JToBI) intonation-labeling scheme [8] provides us with rich information about the tones and break indices, and, their interactions with segmental sounds. Results shown below are based upon the analysis of about 22 hours of spontaneous speech that we have X-JToBI labeled so far.

Figure 5 shows the relationship between the occurrence of two phrase-final boundary pitch movements (BPM) – i.e., L%H% (rising rendition) and L%HL% (rising-falling rendition)-- and the factors Styl and Spnt. The abscissa stands for the rated values of Styl and Spnt, and the ordinates stand for the percentage of the number of occurrence of these BPMs to the total number of break indices that are equal to or stronger than ordinary accentual phrase (i.e. 2, 2+b, 2+bp, and 3. See [8]).

In Figure 5, the curve of L%H% and that of L%HL% moved toward exactly opposite directions. The correlations with the speaking style and spontaneity are positive and negative in the case of L%H%, but negative and positive in the case of L%HL%. The same inverse relationship can be observed in Figures 6, where occurrence rate of the two BPMs are shown as a function of Sex and Type.

These findings suggest strongly that the two BPMs are associated with different pragmatic, or paralinguistic, meanings as are the cases with many of the segmental and morphological variables examined in the previous section.

Also, it seems to be the case that the two BPMs differ with respect to their function of discourse segmentation. See [9] for the effectiveness of the X-JToBI labels for the analysis of discourse boundary strength. Moreover, there are more BPM categories annotated in the X-JToBI labeling of the CSJ-Core. Analysis of other BPMs will certainly reveal more findings about the variations of prosody.

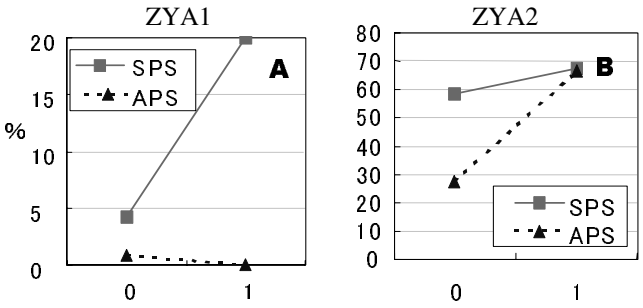

Figure 4. Interaction between Laugh (abscissa) and Type in the analysis of ZYA. The ordinate stands for the coalescence rate [%].Panel A and B stand respectively for coalescence of particle (ZYA1) and auxiliary verb (ZYA2).

5. CONCLUSION

ISBN 1-876346-48-5

The current study examined segmental, morphological, and prosodic variations recorded in the *Corpus of Spontaneous Japanese*. All variables showed clear correlations with more than a single factor extracted from the CSJ.

As a whole, this study revealed that CSJ could be an excellent research resource for the study of linguistic variation. We believe firmly that the public release of CSJ, which is planed to be in the spring of 2004, would lend a strong impetus to the study of linguistic variation in Japanese.

Acknowledgment: We thank Sadaoki Furui and other colleagues of the "*Spontaneous Speech: Corpus and Processing Technology"* project where CSJ is developed.

REFERENCES

[1] Labov, W. *Sociolinguistic Patterns.* Philadelphia, Univ. Pennsylvania Press, 1972.

[2] Maekawa, K, H. Koiso, S. Furui and H. Isahara. "Spontaneous speech corpus of Japanese". *Proc. 2nd Int. Conf. Language Resources and Evaluation (LREC)*, 2, 947-952, Athens, 2000.

[3] Maekawa, K. "Corpus of Spontaneous Japanese: Its design and evaluation". To appear in *Proc. ISCA&IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, Tokyo, 2003.

[4] Maekawa, K. and H. Kikuchi. "Corpus-based analysis of vowel devoicing in spontaneous Japanese". To appear in J. van de Weijer et al. (Eds), *Voicing in Japanese*. LeidenUniveristy (http://www.let.leidenuniv.nl/ulcl/faculty/vdweijer/jvoice/) .

[5] Maekawa, K. "Shortening of lexical long vowels in spontaneous speech". *Proc. 2002 Spring Meeting of Kokugo Gakkai,* 43-50, 2002.

[6] Maekawa, K. "Study of language variation using Corpus of Spontaneous Japanese". *Journal of Phonetics Society of Japan*, 6-3, 48-59, 2002.

[7] Koiso, H., M. Saito, Y. Mabuchi and K. Maekawa. "Hanashikotobaniokeru joshino hatsuonkagenshouno jittai". *Proc. 10th Convention of Shakaigengo Kagakukai*, 215-220, 2002.

[8] Maekawa, K., H. Kikuchi, Y. Igarashi and J. Venditti. "X-JToBI: An extended J_ToBI for spontaneous speech". *Proc. 7th Int. Cong. Spoken Language Processing (ICSLP)*, 3, 1545-1548, Denver, 2002.

[9] Yoneyama, K., J. Fon and H. Koiso. "Durational and prosodic patterning at discourse boundaries in Japanese spontaneous monologs". *Proc. 15th Int. Cong. Phonetic Sciences* (This volume), 2003.

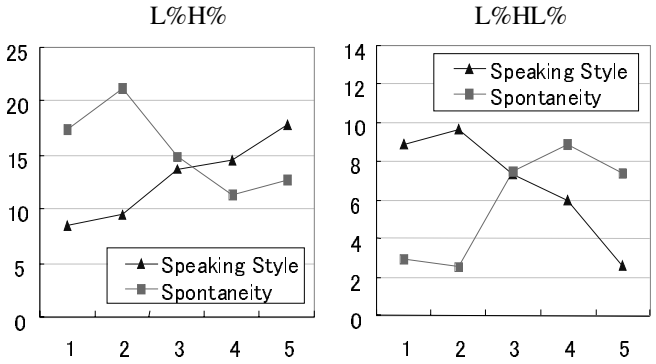

Figure 5. Relationship between the occurrence rates of BPMs [%] and the factors Styl and Spnt (abscissa). See text.

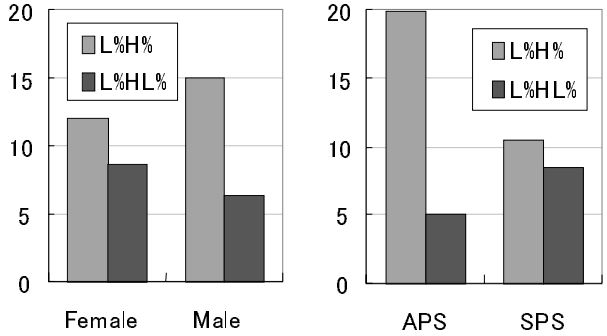

Figure 6. Relationship between the occurrence rates of BPMs [%] and the factors Sex and Type (abscissa). See text.

Table 1. Summary of the results of one-way ANOVA

		FACTORS					
VARIABLES	N	Type	Srate	Styl	Spnt	Laugh	Sex
DV	300,018	****	****	****	****	NS	****
SLV	47,886	****	NS	****	****	****	****
ZYA1 (particle)	1,730	****	NS	****	***	***	NS
ZYA2 (aux. verb)	1,707	****	NS	****	****	****	NS
NO1 (case)	32,317	**	NS	****	**	NS	****
NO2 (nominal)	16,900	****	****	****	****	****	****

Significant at **** P<. 0001, *** P<. 001, ** P<. 01, NS P>=. 01