Design, Compilation, and Some Preliminary Analyses of the Corpus of Spontaneous Japanese

Kikuo Maekawa
National Institute for Japanese Language

1. Introduction

During the last decade or so, we saw a consensus growing among the researchers of spoken language about the need for the study of spontaneous speech. Personally, from a linguist's point of view, I think the study of spontaneous speech has two important aims: to test the validity of experimental and/or theoretical studies using a large amount of data, and, to know the diversity of the usage of a language in its real life.

Achievements of these aims are not easy tasks, however. Most of all, the difficulty lies in the fact that study of spontaneous speech requires large amount of data, because it is difficult to apply experimental design techniques about the content of spontaneous speech (this is a part of the definition of the study of 'spontaneous' study), and, the complexity and diversity of spontaneous speech are sometimes beyond our imagination, at least at the present state of the art.

It seems to me that, even if we delimit the range of observation very narrowly in terms of the genres of spoken language, partial achievement of these aims requires several million word of data that correspond to several hundred hour of speech.

What makes the situation even worse is the compilation cost of spoken corpora: it is by far higher than the cost of written corpora. Even the cost of storing speech signal in a computer was exceedingly too high for linguists until very recently, not to mention the cost of making fine transcription and annotations. All these facts have been preventing linguists and phoneticians from starting full-fledged analyses of spontaneous speech for a long while.

Recently, however, there was also a need for such corpus among the speech engineers working in the field of automatic speech recognition and/or synthesis. In addition, innovation and prevalence of information processing technology, the decline in the cost of software and hardware notably, opened up a possibility to compile a corpus of reliable size within the limit of research budget that was affordable by the joint research group of national laboratories.

It was under these circumstances that we planned the corpus known as *Corpus of Spontaneous Japanese,* or CSJ. We started the compilation of CSJ in the spring of 1999 aiming at the final release in the spring of 2004 (Maekawa et al. 2000).

In this paper, I will outline the design of the corpus, present the status quo of the compilation work, and show the results of preliminary linguistic analyses, concentrating mainly on the analysis of linguistic variations.

2. What is CSJ?

CSJ is a large-scale annotated corpus of spontaneous Japanese. CSJ is an outcome of the national priority-are project known as *Spontaneous Speech: Corpus and Processing Technology* (1999-2003) supported by the Ministry of Education, Culture, Sports, Science and Technology. This is a collaborative work of the National Institute of Japanese Language (NIJL), the Communications Research Laboratory (CRL), and the Tokyo Institute of Technology (TiTech). The project supervisor is professor Sadaoki Furui of TiTech.

2.1. The size and structure of CSJ

2.1.1. The whole corpus

The whole CSJ contains about 660 hours of spontaneous speech that corresponds to about 7 million words (counted as SUW, see 4.2 below). All these speech material is recorded using head-worn close-talking microphone and DAT, and down-sampled to 16kHz, 16bit accuracy. The speech material is transcribed using a two-way transcription scheme designed specially for CSJ. Also, POS (part-of-speech) analysis based upon two different kinds of 'word' is applied for the whole corpus.

2.1.2. The *Core*

There is a true subset of CSJ, called the *Core,* which contains about 500k words or 44 hours of speech. The *Core* is the part of CSJ to which we concentrate the cost of annotation. In addition to the two-way transcription and two-way POS analysis mentioned above, segment label, intonation label, and other miscellaneous annotations are provided for the *Core* (See 4.4-4.6 below).

3. Sources

3.1. APS and SPS

Table 1 shows the sources of spontaneous and other speech material contained in CSJ.

APS (Academic Presentation Speech) is live recording of academic presentation in nine different academic societies held in 1999-2001. The societies range from engineering (3 societies, 621 files), humanities (4 societies, 187 files), and social and behavioral sciences (2 societies, 169 files). Most of the APS are 12-25 minutes long, but there are some hour-long plenary lectures, too.

In addition to these 'genuine' APS, 30 recordings are also counted as APS. These involve hour-long public lectures of archeology or history done in front of large layman audience (6 files), hour-long classroom lectures of Japanese linguistics in front of layman audience (10 files), and academic talks done in closed informal academic meetings in the NIJL (14 files).

SPS (Simulated Public Speaking) is layman's "speech" on everyday topic of about 10-12 minutes in front of small friendly audience. Speakers of SPS are paid layman subjects balanced both in their sex and age. 50 speakers (5 males and 5 females for each of 5 different age groups ranging from their twenties to sixties or older) made a group and were given a set of topics. A set consists of three broad topics.

Table 1: Sources of the Corpus of Spontaneous Japanese.

Sources	Number of Speakers	Number of files	Type	Spontaneity	Time [h]
APS	838	1007	Monologue	Spontaneous	299.5
SPS	580	1683	Monologue	Spontaneous	324.1
Reading text	*(244)	491	Monologue	Read	14.1
SPS by interviewee	*(16)	16	Monologue	Spontaneous	3.4
Interview on APS	* (10)	10	Dialogue	Spontaneous	2.1
Interview on SPS	* (16)	16	Dialogue	Spontaneous	3.4
Task oriented dialogue	* (16)	16	Dialogue	Spontaneous	3.1
Free dialogue	* (16)	16	Dialogue	Spontaneous	3.6
Reproduced APS	* (16)	16	Monologue	Read	5.5
				Total time in hour	658.8

* Parenthesized speakers are already counted in APS or SPS.

Table 2 shows the topics used in the recording. Speakers were given a set of topics about 48 hours prior to the recording time, and asked to prepare three independent talks. Prepared-to-be-text was not allowed. But it was encouraged to make an outline of their talks. As for the speakers recorded in the first year, no broad topic was specified. These speakers spoke about one or two topics that they chose freely.

Table 2: Topics given to the speakers of SPS.

#	Broad topic	Number of files
0	(Not specified)	222
1	Joyful memory of my life	137
2	Sad memory of my life	134
3	The town I live in	134
4	This is what I'm interested in.	151
5	Impressive event of my life	167
6	Commentary on recent news	152
7	If I go to an isolated island, I will bring …	101
8	How to make …	151
9	History of …	100
10	My most precious thing/people	100
11	Things that I want to endow for the 21st century.	150

Most of the SPS recording was done in the recording studio of the NIJLA using the same equipments as in the APS recording. But the recording of the first 50 SPS speakers were outsourced and done in a non-soundproof room. Although the same recording equipments were used, the room acoustics was different.

3.2. Other sources

APS and SPS are the two main sources of CSJ. In addition to these two, we also made recording of speech materials of different nature.

In table 1, "SPS by interviewee" means, literally, SPS spoken by those speakers who took part in the dialogue recording explained below. "Reading text" means reading of two written passages excerpted from two popular science books. 487 SPS speakers read these passages. Time required for reading was 3-4 minutes.

"Interview on APS" is the interview with an APS speaker about the content of his/her APS. The interviewees are those speakers of "SPS by APS speakers". "Interview on SPS" is the interview with the same speakers of "Interview on APS" about the content of their SPS. These interviews were 10-15 minutes long.

The same pairs of interviewee and interviewer participated in "Task oriented dialogue" also, in which they were asked to determine the rank of supposed-to-be performance fee of 8 TV personalities. The same pair of speakers participated in "Free dialogue", also.

Lastly, "Reproduced APS" is the reading of transcribed APS by the same original speaker. The speakers were asked to read aloud the transcribed text of their talks 'faithfully,' i.e., without removing filled-pauses and other disfluencies transcribed in the text.

4. Annotations

4.1. Transcription and tag

Recorded speech is transcribed in two different ways: orthographic and phonetic transcriptions (**小磯・土屋・間淵・斉藤・籠宮・菊地・前川** 2001). In "orthographic" transcription, speech is transcribed using Kanji (Chinese logograph) and Kana (Japanese syllabary) like ordinary Japanese text, but unlike the ordinary writing, our orthographic transcription has rigorous rules of the usage of Kanji and Kana letters. In ordinary text, for example, there are more than five ways of transcribing the phonemic string of /hanasiai/ ("meeting") using Kanji and Kana, but in our orthographic transcription, only one is allowed. Orthographic transcription is useful in making queries of text.

"Phonetic" transcription is written exclusively in Kana letters so that the phonetic details of the utterance being transcribed can be traced. Phonetic transcription is needed for two main purposes. For one, it specifies the reading of many Kanji strings that have multiple readings. For another, it is useful for the study of phonetic/phonological variations of spontaneous speech. Table 3 shows the tags inserted in the transcription text.

Table 3: Tags used in the transcription of CSJ.

Type I: Tags that refer to the characteristics of linguistic message		
Tag	Usage	Example
(D)	Word fragment, repairs	(D ko) korewa
(W)	Reduced or incorrect pronunciation	(W midair; hidari)**
(?)	Uncertainty	(? taonguH)
(F)	Filled pauses	(? anoH, aNnoH)
(M)	Meta-linguistic expression	(M wa) wa (M ha) to kaku
(O)	Foreign language, archaic Japanese etc.	(O that's fine)
(R)	Named entity	(R koiso) san ga
(A)	Use of alphabets in orthographic transcription	(A iHuH; EU)**
(K)	Exceptional use of Kana in orthographic transcription	(K taci (F N) bana; XX) XX stands for Kanji letter of /tacibana/
(S)	Out-of-dictionary items (casual expression)	(S korya)
(Laugh)* (Cry) (Cough) (Yawn)	Speaking while laughing Speaking while crying Speaking while coughing Speaking while yawning	(Laugh nani sore)
(L)	Whispery voice	
Type II: Tags that refer to the existence of phonetic / non-verbal events		
<H>	Non-lexical lengthening of vowels	sorede<H> [sorede:]
<Q>	Non-lexical lengthening of consonants	kai<Q>seki [kais:eki]
<FV>	Vowel that is unable to identify its phonemic status	sorede<FV>
<Breath>* <Laugh> <Cry> <Cough>	Breathing noise Laughter (not speaking) Cry (not speaking) Cough (not speaking)	aru wake desu <Breath>

* The tag letters in this cell are Kanji.

** Entity to the left of ';' is incorrect. Entity to the right is supposed-to-be correct form.

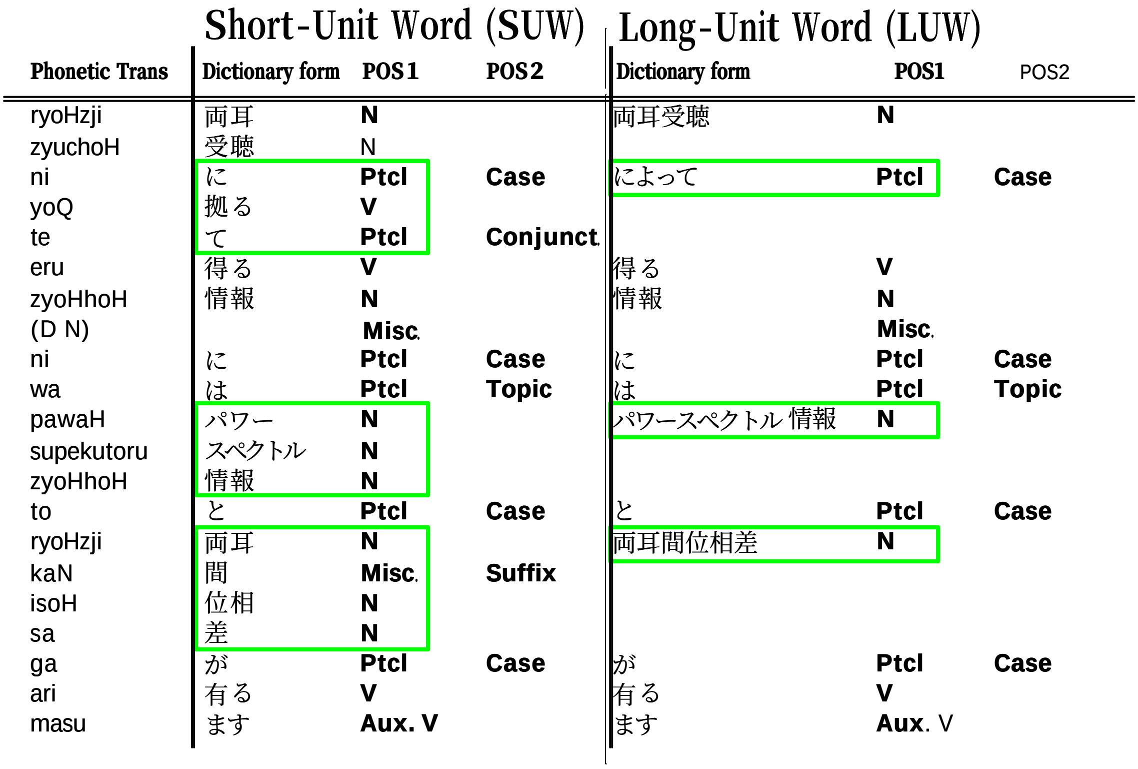
*** Entity to the left of ';' is written in Kana. Entity to the right is in alphabet.

4.2. POS information

Two different POS systems were prepared for CSJ(? ?   2001): short-unit word (SUW) and long-unit word (LUW). Most of the SUW are mono-morphemic words or words made up of two consecutive morphemes, and approximate dictionary items of ordinary Japanese dictionaries.

LUW, on the other hand, is for compounds. For example, the Japanese name for NIJL /kokuricukokugokenkyuHzyo/ consists of one LUW and is analyzed into four SUW's, namely, /kokuricu/ ('national'), /kokugo/ ('Japanese'), /kenkyuH/ ('research'), and, /zyo/ ('institution').

Although most LUW are the combination of consecutive noun SUW or verb SUW, there are also cases where combination of particles or combination of a particle and a verb makes up a LUW. For example, a particle LUW /niyoQte/ ('by') is the combination of a particle /ni/ (place particle) followed by a verb /yoru/ ('due to'), which in turn is followed by another particle /te/ (conjunction particle). Figure 1 compares the SUW and LUW analyses of a short APS sample. The text is a short excerpt from an APS about speech perception. Areas enclosed by rectangles are analyzed differently. POS1 and POS2 are two top levels of hierarchical POS classification.



	Short-Unit Word (SUW)			Long-Unit Word (LUW)		
Phonetic Trans	Dictionary form	POS1	POS2	Dictionary form	POS1	POS2
ryoHzji	両耳	N		両耳受聴	N	
zyuchoH	受聴	N				
ni	に	Ptcl	Case	によって	Ptcl	Case
yoQ	拠る	V				
te	て	Ptcl	Conjunct.			
eru	得る	V		得る	V	
zyoHhoH	情報	N		情報	N	
(D N)		Misc.			Misc.	
ni	に	Ptcl	Case	に	Ptcl	Case
wa	は	Ptcl	Topic	は	Ptcl	Topic
pawaH	パワー	N		パワースペクトル 情報	N	
supekutoru	スペクトル	N				
zyoHhoH	情報	N				
to	と	Ptcl	Case	と	Ptcl	Case
ryoHzji	両耳	N		両耳間位相差	N	
kaN	間	Misc.	Suffix			
isoH	位相	N				
sa	差	N				
ga	が	Ptcl	Case	が	Ptcl	Case
ari	有る	V		有る	V	
masu	ます	Aux. V		ます	Aux. V	

Figure 1: POS analyses based on SUW and LUW.

4.3. Impressionistic rating

During the course of recording, one of recording staffs evaluated subjectively the way the speech being recorded was spoken. This is necessary because the speaking style and spontaneity of speech differ considerably from talk to talk even within a single category of speech type (APS for example).

Five-scale rating was used for the evaluation of the following characteristics: 1) spontaneity, 2) field-specific technical terms, 3) speaking rate, 4) clearness of speech articulation, 5) dialectal features (in terms of lexicon, segmental sound, and lexical accent), and, 6) speaking style.

Our preliminary analyses shown below revealed the effectiveness of the impressionistic rating in the analysis of linguistic variation.

4.4. Segment labeling

The Core of the CSJ is segment labeled. The labels are basically phonemic, but some phonetic labels are used, too. Phonetic labels are needed for the study of phonetic variations.

The phonemic labels were generated automatically from the phonetic transcription, and were aligned automatically to speech signals using the Hidden Markov Model-based algorithm. Human experts, then, adjusted the location and category of automatically generated labels. See the paper by Kikuchi and Maekawa in this volume, and Kikuchi and Maekawa (2003) for the details of the segment labeling techniques and the evaluation of the labeled data.

4.5. Intonation labeling

The Core is also intonation labeled. Because the intonation of spontaneous Japanese differed considerably from that of the read speech, we newly proposed an extended version of the traditional J_ToBI scheme called X-JToBI (Maekawa, Kikuchi, Igarashi and Venditti, 2002). In the new X-JToBI scheme both the tone and BI labels were considerably extended to match the diverse variation of spontaneous speech intonation. Here are some examples of the X-JToBI extensions.

- Enlargement of the phrase-final boundary tone: Two new boundary pitch movement (BPM) categories, i.e., L%LH%, L%HLH%, were added to the traditional two, i.e., L%H% and L%HL%.
- Time-decomposition of complex boundary tones: Complex boundary tones like L%HL% are decomposed into its constituent labels, L%, pH, and HL% in this case, and each of the constituent labels has it is own time stamp. The 'pH' label is called "pointer" and used to denote the peak in the F0 contour. Similarly, L%LH% is decomposed into L%, pL, and LH%. In this way, exact match between the tone labels and their corresponding F0 event was assured.
- Enlargement of the BI inventory: New BI labels were introduced to denote prosodic boundaries caused by the occurrence of various disfluency phenomena like filled-pauses, fragmented word, word-internal pause. BI labels were also enlarged so that it could represent intermediate strength of the prosodic boundary. Foe example, the labels like "2+b" and "2+bp" stand for the prosodic boundary whose strength is in between the traditional "2" and "3". The alphabets in these labels denote the "reasons" why the strength of the boundary in question was judged to be intermediate. For example, "2+b" is used for the cases where an accentual phrase is ended with a BPM, not followed by a pause, and, pitch range resetting was not observed across the boundary in question. Similarly, "2+bp" is applied for the cases where an accentual phrase is ended with a BPM, followed by a pause, and, pitch range resetting was not observed.

The characteristics of the X-JToBI scheme are discussed more closely in Maekawa et al. (2002). See also the paper by Kikuchi and Maekawa in this volume.

4.6. Miscellaneous annotations

The annotations explained so far were all in the blueprint of CSJ. Prompt progress of the compilation, however, made it possible to add some new annotations to the Core. The new annotations include 1) clause boundary (Takanashi, Maruyama, Uchimoto, and Isahara, 2003), 2) discourse segment boundary (Takanashi, Maruyama, Uchimoto, and Isahara., 2003), and, 3) dependency structure within a clause. All these annotation works are currently underway at CRL.

5. Preliminary analyses

The full-fledged evaluation of the CSJ has not been conducted, but here are the results of some preliminary investigations showing hitherto unknown characteristics of spontaneous speech (Maekawa, 2003; Maekawa, Koiso, Kikuchi and Yoneyama, 2003).

5.1. Speaking rate

Figure 2 compares the speaking rates (number of morae per sec) of CSJ and ATR read-speech database (Sagisaka, Takeda, Abe, Katagiri, Umeda, and Kuwabara, 1990). The data for CSJ is computed from the SPS samples in the Core. The data for read speech is computed using the reading of ATR-503 sentences by six male speakers.

This figure shows that, in addition to the higher average speaking rate (8.01 as opposed to 7.11 of ATR), CSJ is characterized by its greater standard deviation (2.07 as opposed to 0.96 of ATR).

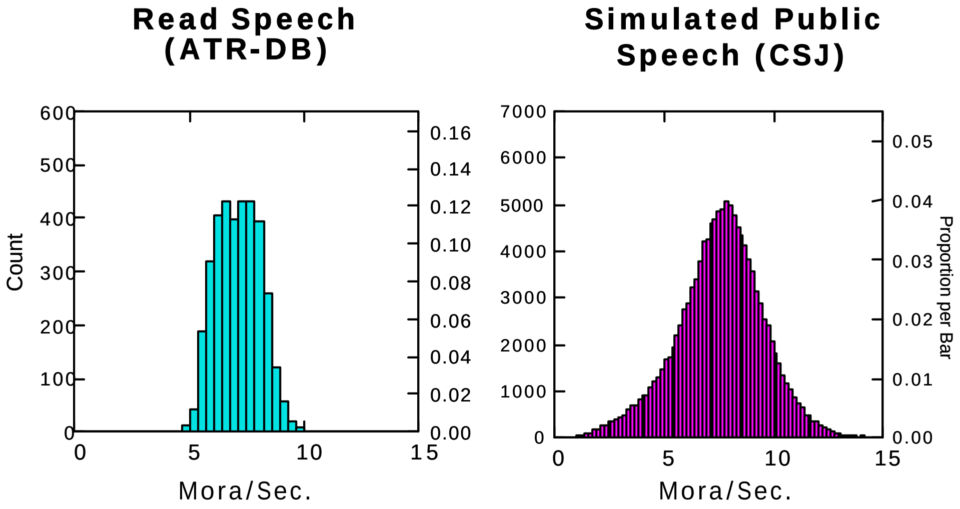


Figure 2: Comparison of speaking rate in read and spontaneous speech.

Figure 3 compares the speaking rates of the ATR-503 sentences, SPS of CSJ, and two different APS’s in CSJ. SPS is faster than read-speech (ATR), but APS is even faster. Also, it is interesting to see that APS of engineering societies are faster than that of humanities. This inter-APS difference may be due to the difference of time allotted to the talkers. Generally speaking, the presentation time in the humanity society (25-35 min.) is longer than that of engineering (15-20 min.).

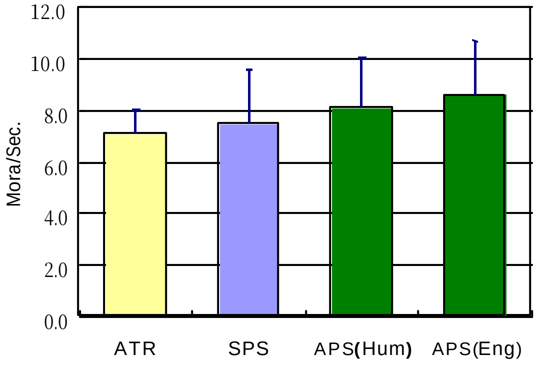


Figure 3: Comparison of speaking rate between SPS and APS of CSJ.

5.2. Disfluency

Disfluency is the most salient feature of spontaneous speech. Among the tags used in the transcription of CSJ, three tags are deeply concerned with disfluency. Tag '(D)' marks word fragment (more exactly, fragment of SUW), tag '(W)' marks reduced or incorrect pronunciation, and, tag '(F)' marks filled-pauses. Figure 4 shows the ratio of these tags to the total number of SUW as a function of speech type (APS vs. SPS) and speakers' sex. Males had more disfluency than females regardless of the speech type. At the same time, SPS contained more disfluency than APS as long as (D) and (W) are concerned. It is interesting to see that (F), the most frequent disfluency, behaved differently from (D) and (W). APS contained more filled-pauses than SPS. Since, pretheoretically, APS is supposed to be relatively less spontaneous than SPS, this casts doubt on the belief that filled-pauses are the good indicator of speech spontaneity.

Figure 5 examines how the ratios of (D), (W), and (F) are correlated with the impressionistic rating of spontaneity. Although all (D), (W), and (F) correlate positively with judged spontaneity (1 and 5 being the least and most spontaneous), correlation of (F) is less linear than those of (D) and (W). As for (F), significant difference exists only between 1 and 2. This fact suggests that (F) could be a good indicator of dichotomy between the read and spontaneous speech, but is not a good indicator of the degree of spontaneity.

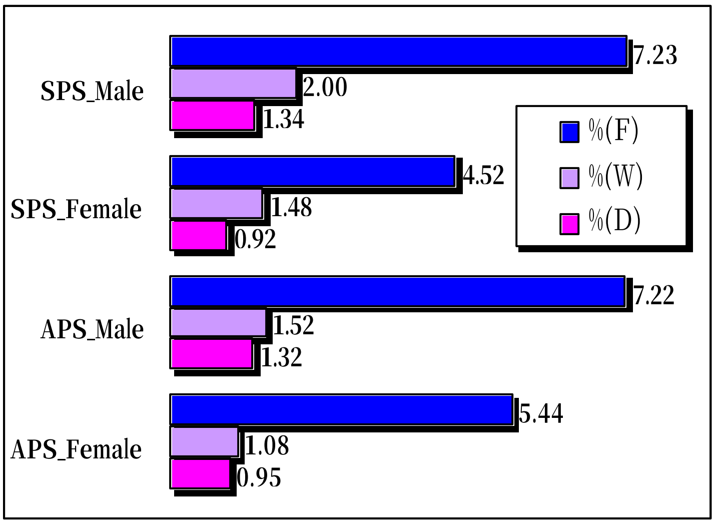


Figure 4: Frequency of (F), (W), and (D) tags in CSJ.

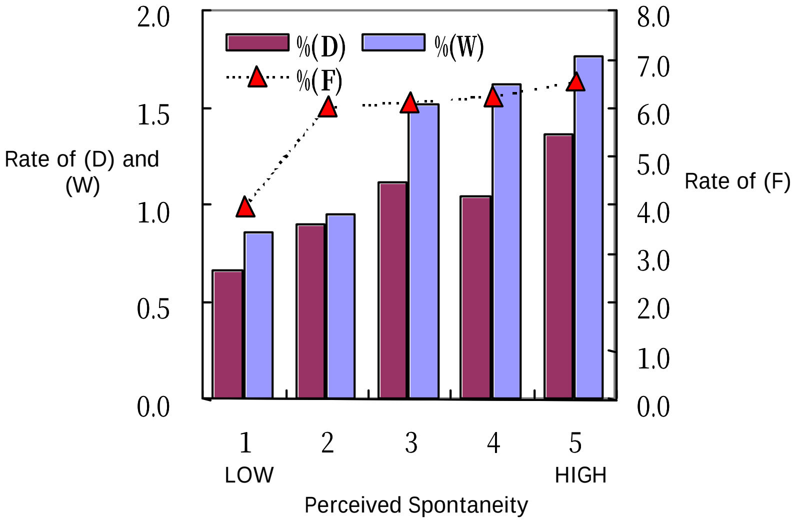


Figure 5: Correlation of disfluency tags and impressionistic rating of speech spontaneity. Left ordinate stands for %(D) and %(W), and, right ordinate for %(F).

5.3. Vowel devoicing

In Japanese, close vowels, /i/ and /u/, tend to be devoiced when they are preceded and followed both by voiceless consonants. This tendency is especially clear in Tokyo Japanese, which is the body of Standard Japanese. Some phonemic analyses, accordingly, describe devoiced vowels as the conditional variants of voiced vowels. However, corpus-based analysis reveals that this is too much a simplification (Maekawa and Kikuchi, forthcoming).

Table 4 shows the effect of the manners of adjacent consonants. This is based upon the analysis of 300,018 segment-labeled vowels in the Core (as of August 2002). Numbers in each cell shows the devoicing rate [%]. The following fricative significantly lowers devoicing rate. This is presumably because vowel devoicing in this context results in succession of two frication noises whose boundary is very difficult to perceive.

Table 4: Effect of the manner of adjacent consonants on the rate [%] of close vowel devoicing.

		FOLLOWING CONSONANT		
VOWEL	PRECEDING CONSONANT	Affricate	Fricative	Stop
/i/	Affricate	81.1	33.3	89.4
	Fricative	96.3	38.1	98.4
	Stop	80.2	51.5	89.3
/u/	Affricate	77.2	48.1	94.5
	Fricative	95.1	61.2	97.5
	Stop	80.8	74.0	80.1

Another devoicing-preventing factor is the context of consecutive devoicing: contexts where more than two successive close vowels are all preceded and followed by voiceless consonants. Figure 6 examines the cases where two close vowels are in the context of consecutive devoicing. Abscissa of the figure is the combination of the consonant manner of the first and second morae containing close vowels. ‘F’, ‘A’, and ‘S’ stand respectively for fricative, affricate, and stop.

Here again, combination of consonant manners plays a crucial role. Devoicing rate of the vowels in the first mora remains high as long as the consonants of the second mora begin with a burst sound (i.e. either stop or affricate), while the rate is low in cases where the consonants of the second mora begin frication noise (i.e. fricative).

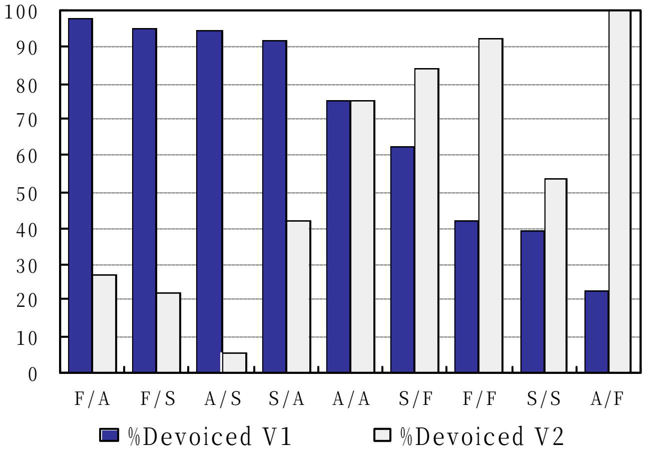


Figure 6: Effect of the combination of the manner of consonants upon the rate [%] of consecutive devoicing.

Figure 7 shows the influence of speaking rate upon close vowel devoicing. Abscissa in this figure stands for speaking rate normalized within a speaker. Speaking rate [mora/sec] is computed for all pause-separated utterances, and the slowest and fastest 25% are classified as 1 and 4 respectively. Devoicing rate is positively correlated with the speaking rate. Lastly, figure 8 shows that speaking rate has similar effect upon the devoicing of non-close vowels, with the probable exception of /o/.

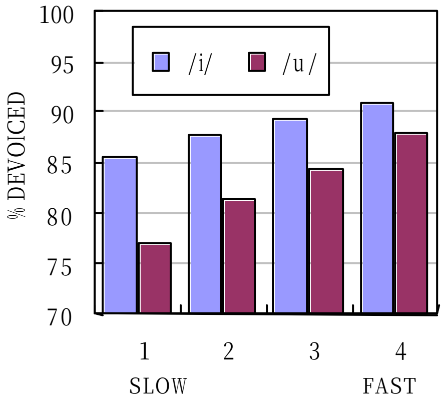


Figure 7: Effect of speaking rate upon close vowel devoicing.

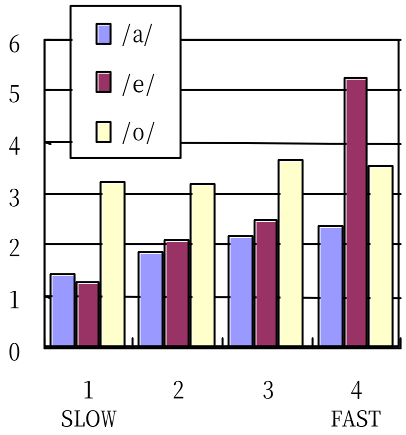


Figure 8: Effect of speaking rate upon non-close vowel devoicing.

5.4. Shortening of lexical long vowels

All five Japanese vowels have the phonological contrast of vowel length like /a/ (short) and /aH/ (long). Contrast between the minimal pairs like /kodoku/ ('solitude') vs. /koHdoku/ ('subscribe'), or, /obasaN/ ('aunt') vs. /obaHsaN/ ('grand ma') are all lexically specified. Sometimes, however, phonological long vowels could be realized as if they were short vowels. For example, /deHtaH/ (English loan of 'data') and /hoNtoH/ (Sino-Japanese meaning 'really') could be realized as /deHta/ and /hoNto/ respectively.

Although the factors governing the shortening have not been fully understood, there are beliefs widely shared by the researchers of the Japanese language: 1) Loan words from European languages (English especially) are shortened more frequently than Sino-Japanese words, 2) High frequency words tend to be shortened, 3) Shortening occurs mostly in word-final position, 4) Shortening tends to occur when a long vowel is followed by another long vowel, 5) Shortening is a characteristic of low speaking-style.

Adequacy of these beliefs is examined using a subset of CSJ, namely, manually POS analyzed 884k SUW sub-corpus (? ? 2002a)

To begin with, the effect of word-class is examined. Table 5 shows that the effect of word-class is statistically significant (P<.0001). But, different conclusion could be obtained when we look at the words with high shortening rate. Figure 9 shows the shortening rate [%] of 30 words that showed the highest shortening rates. Filled and open bars correspond respectively to Sino-Japanese and English-loan. As can be seen from the figure, the word that showed the highest shortening rate (/hoNtoH/ “true”) was a Sino-Japanese, and, 15 out of the total of 30 words were Sino-Japanese.

Table 5: Effect of word-class upon the shortening of lexical long vowels.

WORD CLASS	NOT SHORTENED	SHORTENED	%SHORTENED
English loan	5934	543	8.38%
Sino-Japanese	47665	1004	2.06%

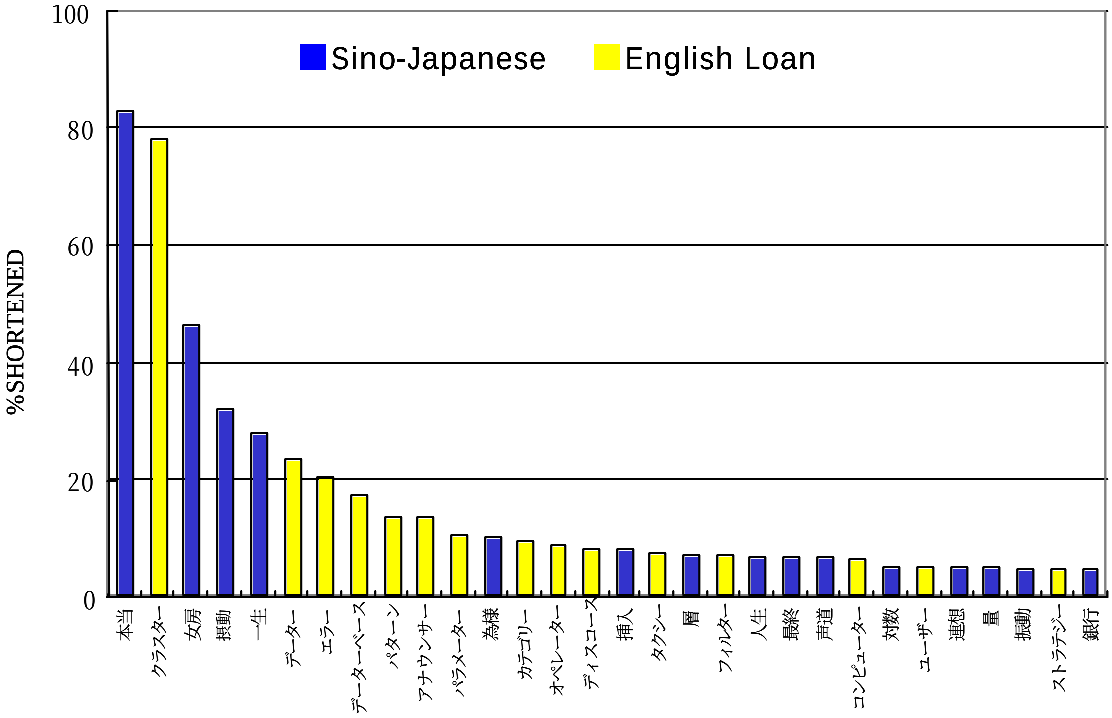


Figure 9: Words with highest shortening rate and their word-class.

Figure 10 is a scatter plot of the word-frequencies in CSJ and the shortening rates. There is no statistically significant correlation (Pearson correlation coefficient was 0.133). If we remove words with higher than 20% shortening rate (whose word-form are shown in the figure), the coefficient becomes -0.018, but this is not significant either. Significant correlation was observed when we remove words with zero shortening rate and taking the logarithm of both axes (r=0.391). Probably, the effect of word frequency is active only with words that tend to be shortened.

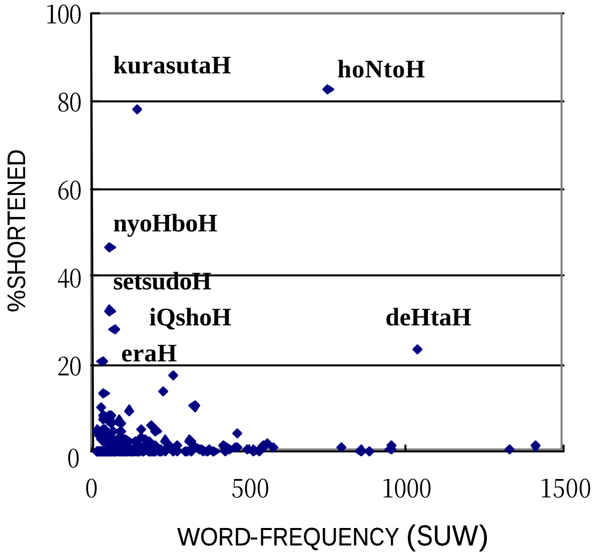


Figure 10: Correlation between the word-frequency and the shortening rate.

Table 6 examines the effect of position in a SUW. Shortening rate varies significantly depending on the position in a SUW. The belief that word-final position is the most enhancing position is supported by this result, but it is the shortening-preventing effect of the word-initial position that is the most remarkable.

Table 7 examines the effect of so-called special morae on shortening. When the long vowel in question is preceded by a special mora, shortening rate increased significantly, but the effect differs considerable depending on the preceding mora. It was moraic nasal that showed the strongest effect as long as this table is concerned.

Table 8 compares the shortening rates between APS and SPS. As expected, shortening rate is higher in SPS where speaking style is expected to be lower than APS. At the same time, however, shortening rate is not negligible in APS.

Table 6: Effect of the position in a SUW upon shortening.

POSITION	NOT SHORTENED	SHORTENED	%SHORTENED
INITIAL	22429	21	0.09%
INTERMEDIATE	2118	83	3.77%
FINAL	29187	1444	4.71%

Table 7: Effect of the preceding special morae upon shortening.

SPECIAL MORAE	NOT SHORTENED	SHORTENED	%SHORTENED
Long vowel /H/	4,917	375	7.09%
Moraic nasal /N/	6,390	699	9.86%
Geminate /Q/	1,647	46	2.72%
No special mora	40,780	428	1.04%

Table 8: Effect of the speech type upon shortening.

SPEECH TYPE	NOT SHORTENED	SHORTENED	%SHORTENED
APS	36,311	907	2.44%
SPS	10,030	637	5.97%

Figure 11 shows the relationships between the shortening rate and the impressionistic ratings of speaking style and speech spontaneity. The abscissa of the figure stands for the judged speaking style and spontaneity, and the ordinate stands for shortening rate. Note left and right ordinate stand respectively for spontaneity and speaking style. Shortening rate decreased monotonically as a function of the increase in judged speaking style, and, increased as a function of the increase in judged spontaneity.

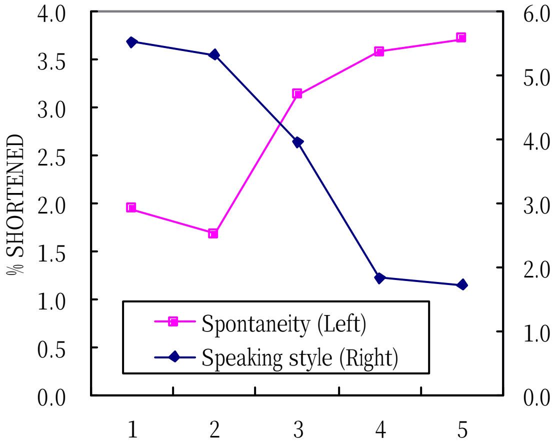


Figure 11: Relationship between the impressionistic ratings (speaking style and speech spontaneity) and shortening rate.

Figure 12 shows the effect of extra-linguistic factor, ‘laughter’. Here, the data is classified according to whether the utterance unit containing the long vowel in question involves the tag ‘(Laugh)’ or not. This figure shows that presence of ‘laughter’ enhances the shortening rate strongly in SPS, but not in APS. Speaking statistically, there is an interaction between the effect of ‘laughter’ and that of speech type.

Lastly, figure 13 shows the interaction between the speaker’s sex and speech type. In APS there was no statistically significant difference between males and female, but in SPS females’ shortening rate was more than tripled compared to that of APS.

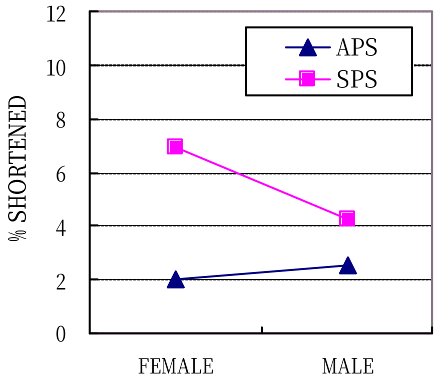


Figure 12: Interaction between 'laughter' and speech type.

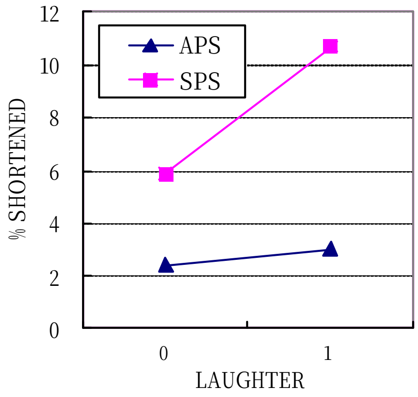


Figure 13: Interaction between speakers' sex and speech type.

5.5. Word coalescence

Word coalescence is a morphological process whereby (more than) two words are fused into one word. Here, we examine the case where /de/ and /wa/ are fused into /zya/ (? ? 2002b). There are two types of /de/+/wa/ sequence differing in the POS of /de/ (/wa/ is always topic particle) : case particle of 'place' (e.g. *Tokyo-de-wa ame-ga futta* "It rained in Tokyo"), and, *rentaikei* of auxiliary verb (or copula) /da/ (e.g. *ame-ga futta-no-wa Tokyo-de-wa nai* "It was not Tokyo that it rained"). Both of these sequences can be coalesced into /zya/ (*Tokyo-zya ame-ga futta*, and, *ame-ga futta-no-wa Tokyo-zya nai*), but the probability of coalescence differs considerably depending on the POS of /de/. Table 9 shows the coalescence rate as a function of POS of /de/ and speech type. Coalescence rate is much higher in auxiliary verb and in SPS.

Table 10 examines the effects of extra-linguistic factor, 'laughter'. Here, the data is classified according to whether the utterance in which the /de/+/wa/ sequence occurred contained the tag '(Laugh)' or not. The co-occurrence with '(Laugh)' enhanced considerably the coalescence. The same enhancing effect of '(Laughter)' was observed in the shortening of lexical long vowel (? ? 2002a).

Table 9: Coalescence rate as a function of POS of /de/ and speech type.

POS OF /de/	SPEECH TYPE	N. /de wa/	N. /zya/	%COALESCED
Case particle	APS	1,311	11	0.8%
	SPS	389	19	4.7%
Auxiliary verb	APS	653	256	28.2%
	SPS	327	471	59.0%

Table 10: Coalescence rate as a function of POS of /de/ and presence of 'laughter.'

POS OF /de/	LAUGHTER	N. /de wa/	N. /zya/	%COALESCED
Case particle	Without	1,686	28	1.6%
	With	14	2	12.5%
Auxiliary verb	Without	956	678	41.5%
	With	24	49	67.1%

Figure 14 shows the relationship between two types of impressionistic rating, spontaneity and speaking style, and the rate of coalescence. Coalescence correlated positively and negatively with spontaneity and speaking style respectively, regardless of the POS difference.

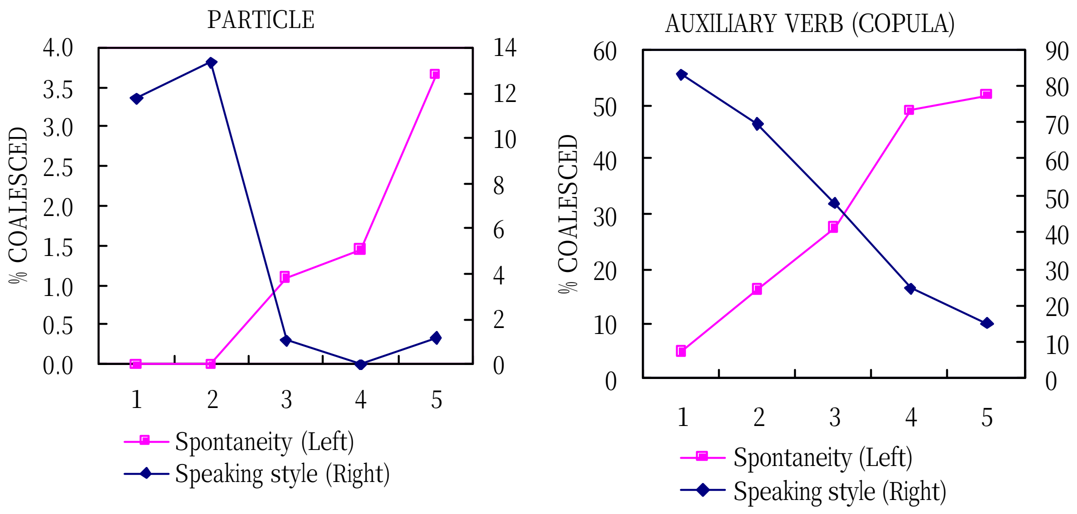


Figure 14: Relationship between the impressionistic ratings (speaking style and speech spontaneity) and coalescence rate.

5.6. Alternation of particle /no/ and moraic nasal

Particle /no/ is pronounced as a single moraic nasal /N/ sometimes. This variation is analyzed as another example of morphological variation (小磯・斎藤・間淵・前川 2002). As was the case in the analysis word coalescence, POS difference plays important role in the alternation, but this time the difference is concerned with sub-classification of particle. There are two types of /no/ particle differing in its grammatical meaning: genitive case particle and nominalization particle

Example of genitive is /hoN wa cukue no ue da/ (‘The book is on the desktop”), where two nouns /cukue/ (“desk”) and /ue/ (“top”) are connected by the particle. Example of nominalization is /kare ga toHkyoH ni iQta no desu/ (“It happened that he went to Tokyo”), where a whole sentence /kare ga toHkyoH ni iQta/ (“he went to Tokyo”) is nominalized by the particle and constitutes an predicate with the help of a copula (/desu/).

Table 11 shows quite prominent effect of the POS difference. It is interesting to see that the effect is consistent across speech types. The effect of ‘laughter’ is analyzed in Table 12. Presence of ‘laughter’ enhances the moraic nasalization regardless of the POS difference.

Table 11: Rate of /N/ as a function of POS of /de/ and type of speech.

POS OF /no/	SPEECH TYPE	N. /no/	N. /N/	% /N/
GENITIVE	APS	20598	75	0.4%
	SPS	11473	71	0.6%
NOMINALIZATION	APS	4097	2681	39.6%
	SPS	4077	6045	59.7%

Table 12: Rate of /N/ as a function of POS of /de/ and the presence of 'laughter.'

POS OF /no/	LAUGHTER	N. /no/	N. /N/	% /N/
GENITIVE	Without	31708	142	0.4%
	With	363	4	1.1%
NOMINALIZATION	Without	7975	8305	51.0%
	With	199	421	67.9%

Lastly, Figure 15 shows the relationship between impressionistic ratings and the rate of alternation. Rate of moraic nasal correlated positively and negatively with speaking style and speech spontaneity respectively regardless of the POS difference.

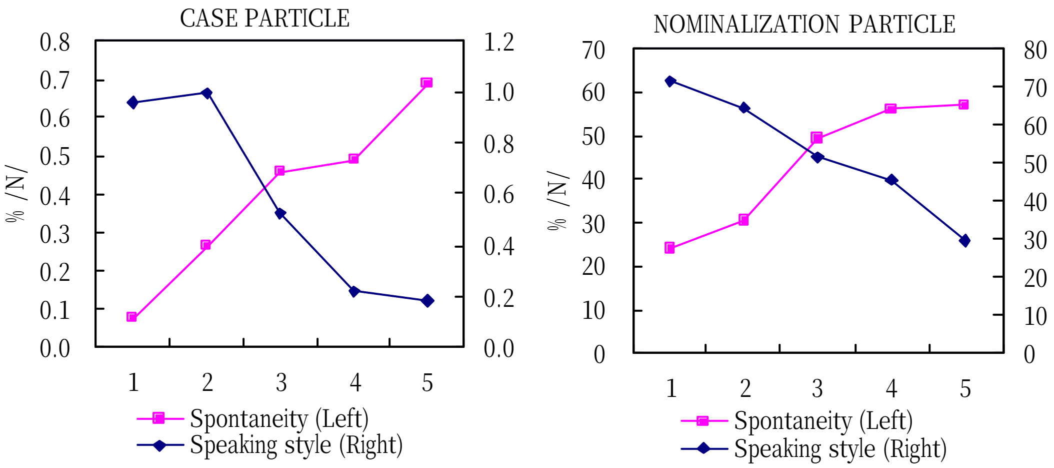


Figure 15: Relationship between the impressionistic ratings (speaking style and speech spontaneity) and the rate of moraic nasal /N/.

5.7. Factors of variation

So far, we have seen that CSJ could be an excellent resource for the study of language variations. Table 11 is the summary of one-way ANOVA applied for the combinations of various linguistic variables and extra-linguistic factors (Maekawa et al. forthcoming).

Among the factors analyzed, 'Type' denotes difference between APS and SPS, 'Spk rate' denotes within-speaker normalized speaking rate. 'Style' and 'Spnt' denote impressionistic ratings of speaking style and spontaneity respectively. 'Laugh' denotes presence vs. absence of '(Laugh)' tag in an utterance, and, 'Sex' denotes speaker's sex.

This tables shows that for all linguistic variables at least 4 factors are statistically significant. At the same time, at least four linguistic variables are influenced by each factor, with the sole exception of the speaking rate.

Incidentally, astute readers might notice that the significance levels shown in the table are all very high; one reason for this is the exceedingly large amount of data used in these statistical tests shown in the second column of the table.

Table 13: Result of one-way ANOVA applied separately for all combinations of linguistic variables and factors of variation.

		FACTORS					
VARIABLES	N	Type	Spk rate	Style	Spnt	Laugh	Sex
Vowel devoicing	300,018	****	****	****	****	NS	****
Shortening of long vowels	47,886	****	NS	****	****	****	****
Coalescence: ZYA1 (particle)	1,730	****	NS	****	***	***	NS
Coalescence: 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

5.8. Boundary pitch movements

Intonation labeling of the *Core* is currently underway, and expected to be finished by the end of July 2003. Phonetic and phonological investigation of the intonational characteristics of spontaneous speech per se beside, the label is expected to be precious research resource for various studies of spontaneous speech including the analysis of discourse structure (**小磯・米山・槙** 2003, Yoneyama et al. 2003).

Here, distributions of two boundary pitch movements (BPM) are compared. BPM is a characteristic change of pitch that marks the end of a prosodic phrase, and occurs in most cases in the last syllable of the phrase. Intonation of spontaneous speech is characterized both by the richness of BPM inventory and the higher rate of BPM occurrence. The current X-JToBI inventory includes: L%H% (rising), L%LH% (sustained low followed by a rise), L%HL% (rising-falling), and, L%HLH% (rising-falling-rising).

The following figures compare the occurrence rates of the two most frequent BPM: L%H% and L%HL%. The numbers in the figures show the rate of BPM computed as the number of occurrence of a given BPM divided by the total number of break indices (BI) stronger than 2. Note in the X-JToBI, there are BI like ‘2+b’, ‘2+p’, and ‘2+bp’ whose strengths are supposed to be intermediate between canonical 2 and 3.

In figure 16, rates of the two BPM’s are shown as a function of speaker’s sex and speech type. It is interesting to see that the two BPM behaved in nearly the opposite way with respect to both of these factors. Males used more L%H% than females, and, females used more L%HL% than males, and, APS had more L%H% than SPS, and, SPS had more L%HL% than APS. These differences are, most probably, the consequence of style difference associated with these BPM’s.

Figure 17 shows the relationship between the occurrence rates of these BPM’s and impressionistic ratings of speaking rate and spontaneity. Here again, the behaviors of the two BPM’s are in opposite directions. The rate of L%H% was correlated positively and negatively with the perceived speaking style and spontaneity, respectively, while the rate of L%HL% was correlated negatively and positively with the perceived speaking style and spontaneity respectively.

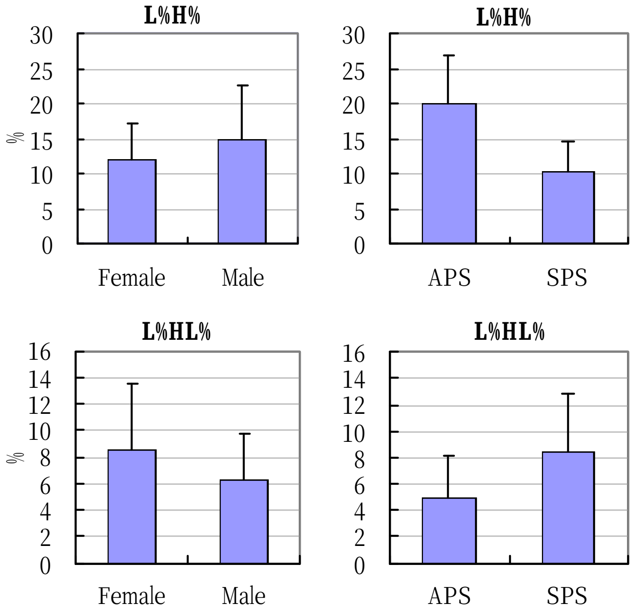


Figure 16: Rates of two BPM's as a function of speaker's sex and speech type.

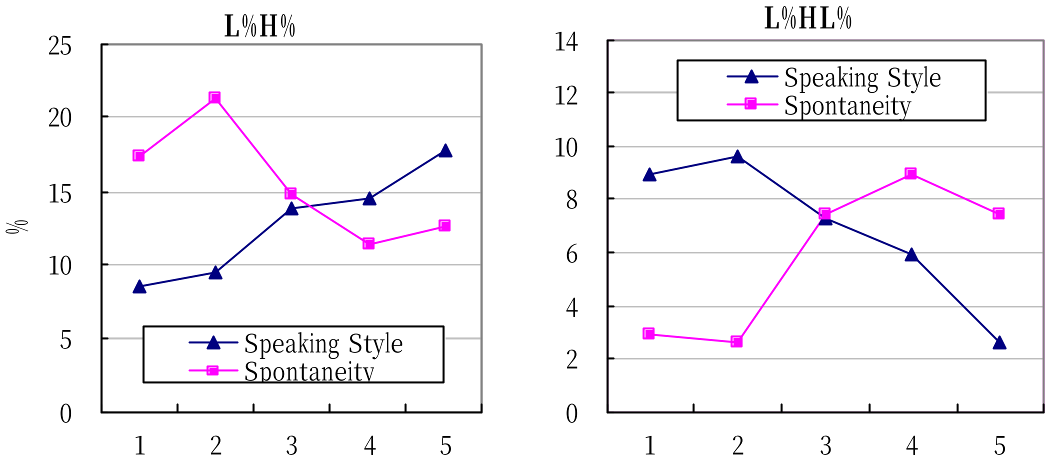


Figure 17: Relationship between the impressionistic ratings (speaking style and speech spontaneity) and the rate of two BPM's.

The distributional difference of BPM's reported above suggest the possibility to make discrimination of speech types simply by knowing the occurrence rates of BPM's. Figure 18 shows that it is the case, at least to some extent.

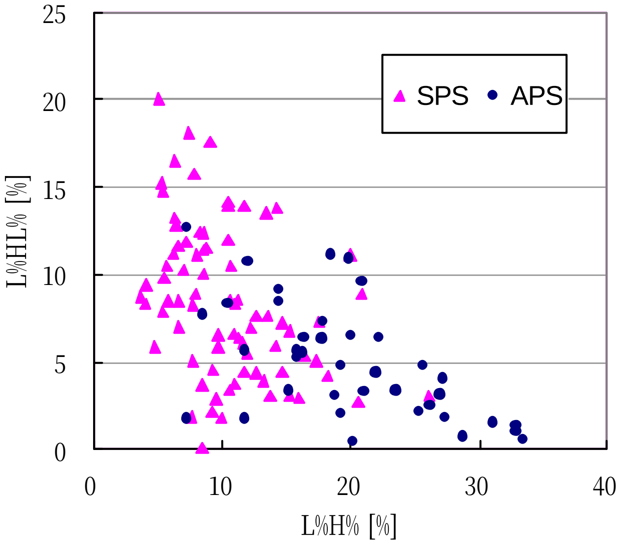


Figure 18: Scatter plot of the occurrence rate [%] of L%H% and L%HL%.

6. Concluding remarks

The last half of this paper was devoted for the analysis of some selected linguistic variations as captured in CSJ, and revealed the usefulness of the corpus as the resource for the study of linguistic variations. But this is definitely not the only area of linguistic research to which CSJ will make contribution. CSJ can provide rich research information for the fields like discourse analysis, psycholinguistics, and speech synthesis, not to mention the study of automatic speech recognition to which CSJ has already made large contribution.

Lastly, it is to be noted that the compilation of CSJ is still underway aiming at the release in the spring of 2004. It is our intention to make the corpus publicly available with reasonable handling fee, for both academic and non-academic users.

Acknowledgments

The author wish to express his gratitude for all speakers of CSJ, who kindly gave us permission to make their talks open for public. He is also grateful for Professor Sadaoki Furui and all other members of the *Spontaneous Speech: Corpus and Processing Technology* project.

References

Kikuchi, H. and Maekawa, K. (2003) Performance of segmental and prosodic labeling of spontaneous speech. *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, Tokyo (pp. 191-194).

Maekawa, K. (2003) Corpus of Spontaneous Japanese: Its design and evaluation. *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, 7-12.

Maekawa, K. and Kikuchi, H. (Forthcoming) Corpus-based analysis of vowel devoicing in spontaneous Japanese –An interim report. To appear in J. van de Weijer, K. Nanjo, and T. Nishihara (eds.) *Japanese Voicing*.

Maekawa, K., Kikuchi, H., Igarashi, Y. and Venditti, J. (2002) X-JToBI: An extended J_ToBI for spontaneous speech.

Proceedings of the 7th International Congress on Spoken Language Processing (ICSLP2002), 3: 1545-1548.

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

Maekawa, K., Koiso, H., Kikuchi, H. and Yoneyama, K. (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)*, Barcelona, 643-646.

Sagisaka, Y., Takeda, K., Abe, M., Katagiri, S., Umeda, T. and Kuwabara, H. (1990) A large-scale Japanese speech database. *Proceedings of the International Congress on Spoken Language Processing*, 1089-1092.

Takanashi, K., Maruyama, T., Uchimoto, K. and Isahara, H. (2003) Identification of "sentence" in spontaneous Japanese. *Proceedings of the ISCA&IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, 183-186.

Takeuchi, K., Takanashi, K., Morimoto, I., Koiso, H. and Isahara, H. (2003) Committee-based discourse purpose assignment: Discourse structure annotations of spontaneous Japanese monologue. *Proceedings of the ISCA&IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, 199-202.

Yoneyama, K., Koiso, H. and Fon, J. (2003) A corpus-based analysis on prosody and discourse structure in Japanese spontaneous monologues. *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, 27-31.

小椋秀樹．(2001)「話し言葉コーパスの単位認定基準について」話し言葉の科学と工学ワークショップ講演予稿集，21-28．

小磯花絵・土屋菜穂子・間淵洋子・斉藤美紀・籠宮隆之・菊地英明・前川喜久雄．（2001）「『日本語話し言葉コーパス』における書き起こしの方法とその基準について」日本語科学，9，43-58．

小磯花絵・斎藤美紀・間淵洋子・前川喜久雄．（2002）話し言葉における助詞の撥音化現象の実態－『日本語話し言葉コーパス』を用いて－，第10回社会言語科学会研究大会予稿集，215-220．

小磯花絵・米山聖子・槙洋一．（2003）『日本語話し言葉コーパス』を用いた談話構造と韻律の関係に関する一考察，人工知能学会研究会資料(SIG-SLUD-A203-P17)，139-144．

前川喜久雄．（2002a）話し言葉における長母音の短呼－『日本語話し言葉コーパス』を用いた音声変異の分析－，国語学会2002年度春季大会要旨集， 43-50．

前川喜久雄．（2002b）『日本語話し言葉コーパス』を用いた言語変異研究，音声研究，6(3)，48-59．