Towards Automatic Transformation between Different Transcription Conventions: Prediction of Intonation Markers from Linguistic and Acoustic Features

Yuichi Ishimoto†, Tomoyuki Tsuchiya†, Hanae Koiso†, Yasuharu Den‡†

†National Institute for Japanese Language and Linguistics
10-2 Midori-cho, Tachikawa, Tokyo 190-8561, Japan
{yishi, ttsuchiya, koiso}@ninjal.ac.jp

‡Faculty of Letters, Chiba University
1-33 Yayoicho, Inage-ku, Chiba 263-8522, Japan
den@cogsci.l.chiba-u.ac.jp

Abstract

Because of the tremendous effort required for recording and transcription, large-scale spoken language corpora have been hardly developed in Japanese, with a notable exception of the *Corpus of Spontaneous Japanese* (CSJ). Various research groups have individually developed conversation corpora in Japanese, but these corpora are transcribed by different conventions and have few annotations in common, and some of them lack fundamental annotations, which are prerequisites for conversation research. To solve this situation by sharing existing conversation corpora that cover diverse styles and settings, we have tried to automatically transform a transcription made by one convention into that made by another convention. Using a conversation corpus transcribed in both the Conversation-Analysis-style (CA-style) and CSJ-style, we analyzed the correspondence between CA's 'intonation markers' and CSJ's 'tone labels,' and constructed a statistical model that converts tone labels into intonation markers with reference to linguistic and acoustic features of the speech. The result showed that there is considerable variance in intonation marking even between trained transcribers. The model predicted with 85% accuracy the presence of the intonation markers, and classified the types of the markers with 72% accuracy.

Keywords: transcription transformation, prediction model, accentual phrase

1. Introduction

There have been lots of attempts to construct large-scale spoken language corpora for the past few decades. Because of the tremendous effort required for recording and transcription, however, large-scale spoken language corpora have not been developed in Japanese, with a notable exception of the *Corpus of Spontaneous Japanese* (CSJ) (Maekawa, 2003). Although CSJ contains a huge amount of monolog speech, such as academic presentation speech and general speech on everyday topics, it contains very few amount of dialog speech, which is the center of our daily linguistic activities. There have been no large-scale conversation corpora in Japanese so far. Although various research groups have individually developed conversation corpora, these corpora are small in size.

The aim of our research project is to solve this situation by sharing existing conversation corpora that cover diverse styles and settings. Although individual corpora so far developed are small, the amount of the data available to the research community will increase dramatically if we share these corpora. These corpora, however, are transcribed by different conventions and have few annotations in common, and some of them lack fundamental annotations such as prosodic information and dialog function, which are prerequisites for conversation research.

As a first step in this endeavor, we are trying to automatically transform a transcription made by one convention into that made by another convention. Our preliminary investigation showed that transcription conventions of Japanese conversation corpora can be classified into two styles: the Conversation-Analysis-style (CA-style) (Jefferson, 2004) and the CSJ-style (Koiso et al., 2006). Using a conversation corpus transcribed in both the CA- and CSJ-styles, we analyze the correspondence between CA's "intonation markers" and CSJ's "tone labels," and construct a statistical model that converts tone labels into intonation markers with reference to linguistic and acoustic features of the speech.

2. Method

2.1. Data

Two dialogs, chiba0232 and chiba0432, from the Chiba Three-Party Conversation Corpus (Den and Enomoto, 2007), which is a collection of casual conversations in Japanese among friends on campus, were used for this study. Each dialog was 10 minutes long, and 6 different speakers participated in the two dialogs. The entire corpus was annotated with utterance units, morphological information, and prosodic information in addition to transcriptions

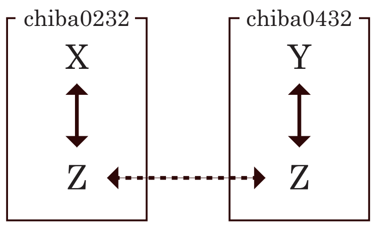

Figure 1: Comparison between transcribers and between data. X, Y, and Z indicate the transcribers.

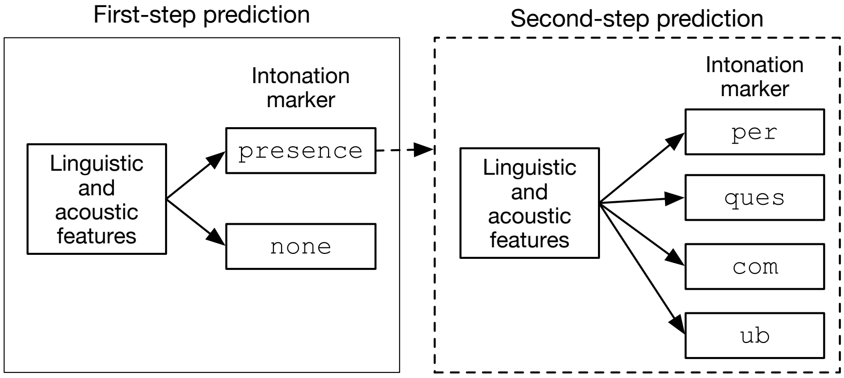

Figure 2: Two-step prediction procedure.

in the CSJ-style (Den et al., 2010).

2.2. Annotation

2.2.1. Tone labels in the CSJ-style transcription

For the CSJ-style transcription, prosodic annotation based on the X–JToBI scheme (Maekawa et al., 2002) is associated; the tone labels for boundary pitch movements are provided at the boundaries at the ends of accentual phrases. The tone labels are either of the following:

L% falling

H% rising

LH% rising with extended low onset

HL% rising-falling

The L% tone does not always indicate an explicit fall in fundamental frequency (F0). It sometimes marks absence of a boundary pitch movement, and, thus, differs from the 'period' in CA's intonation markers described below.

2.2.2. Intonation markers in the CA-style transcription

We focus on the following four intonation markers used in the CA-style transcription.

period (`per`) '`.`' a falling, or final intonation

question mark (`ques`) '`?`' rising intonation

comma (`com`) '`,`' continuing intonation

under bar (`ub`) '_' flat intonation

The CA-style transcriptions were created, for the two dialogs used in this study, by three researchers working in CA: X, Y, and Z. X transcribed chiba0232, Y chiba0432, and Z both of them. All the transcriptions were based on the well-established convention developed by Jefferson (Jefferson, 2004).

The research careers in CA of X, Y, and Z were as follows. X and Z learned CA at the University of California, Los Angeles, and Y at the University of California, Santa Barbara. Each of them had more than six years of experience, including transcribing the data and attending data sessions. Y was also trained in transcription based on Du Bois' convention (Du Bois et al., 1993).

2.3. Analysis and Modeling

First, we analyze data for correspondence between CSJ's tone labels and CA's intonation markers. Next, we examine the variance in intonation marking between transcribers for the same data, as illustrated by the solid arrows in Figure 1. Finally, we construct statistical models to predict CA's intonation markers from CSJ's tone labels as well as linguistic and acoustic features of the speech. Two transcriptions produced by the same transcriber are used for training and testing of the models, respectively, as illustrated by the dashed arrow in Figure 1.

In the statistical modeling, we extracted the following linguistic and acoustic features from each accentual phrase (AP), which were used as predictors of the models.

■Linguistic features

`tone` boundary pitch movement at the end of the AP: L%, H%, HL%, and LH%

`lastPOS` part of speech of the last word in the AP

`penultPOS` part of speech of the penultimate word in the AP

`loc` location of the AP measured by the number of APs counted from the beginning of the utterance

`revLoc` location of the AP measured by the number of APs counted from the end of the utterance

■Acoustic features

`f0MinAP` the minimum F0 value in the AP

`f0MaxAP` the maximum F0 value in the AP

`f0MaxWord` the maximum F0 value in the last word of the AP

`pwrMaxAP` the maximum power value in the AP

`pwrMaxWord` the maximum power value in the last word of the AP

`amdAP` average mora duration of the AP

`lastF0Val` value of the last extracted F0 in the AP

`lastF0Loc` time difference from the point at which `lastF0Val` is extracted to the end of the AP

`lastF0Rise` rising trend of F0 at the end of the AP, which is the margin from the minimum F0 value in the last word of the AP to `lastF0Val`

For prediction models, we used Breiman's random forest algorithm (Breiman, 2001). The prediction of intonation markers was conducted in two steps as described in Figure 2; the first model predicts whether or not an intonation marker is present at the end of an AP, and the second model classifies the type of the intonation marker when the first model detects the presence of any intonation marker.

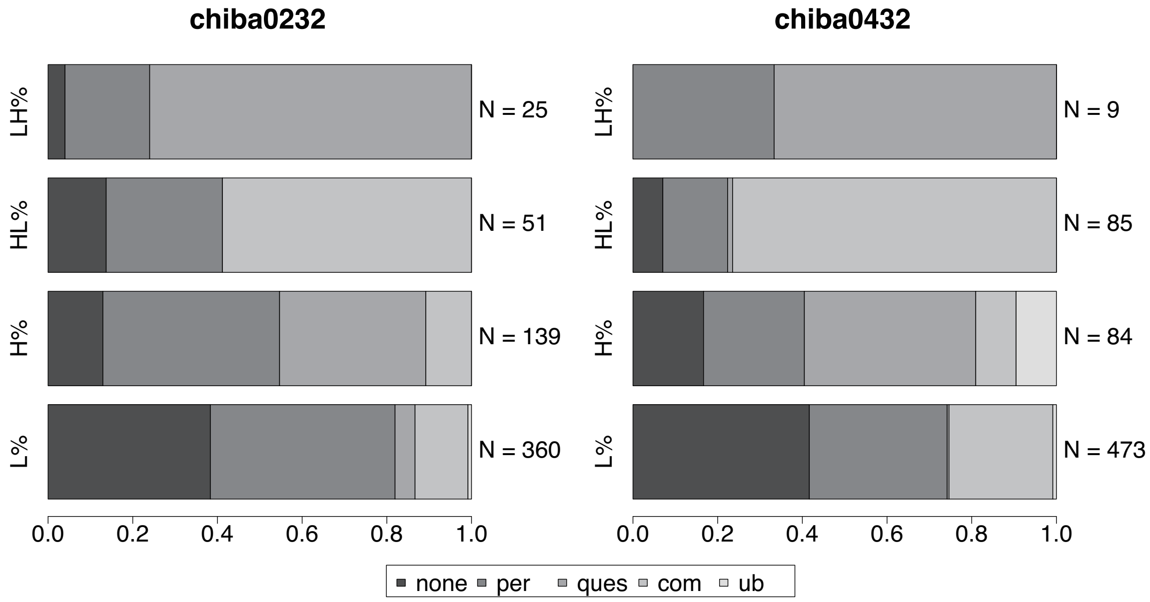

Figure 3: Correspondence between CSJ's tone labels and CA's intonation markers.

Table 1: Variance in intonation marking between transcribers.

chiba0232 (agreement = 76.0%, κ = .66)

	Z				
X	none	per	ques	com	ub
none	**130**	12	4	2	1
per	9	**140**	15	0	0
ques	2	9	**58**	0	1
com	33	20	2	**26**	0
ub	0	1	1	0	**1**

chiba0432 (agreement = 69.9%, κ = .58)

	Z				
Y	none	per	ques	com	ub
none	**184**	8	0	0	0
per	30	**126**	1	3	0
ques	5	11	**29**	1	0
com	92	14	1	**89**	5
ub	0	13	0	0	**0**

3. Results

3.1. Correspondence between CSJ's tone labels and CA's intonation markers

Figure 3 shows the correspondence between CSJ's tone labels and CA's intonation markers. Approximately 40% of the APs with L% labels were unmarked in the CA-style transcriptions, and the remaining 60% were marked as `per` or `com`; the rate of `per` in chiba0232 was higher than that of `com`, while the rates of `per` and `com` in chiba0432 were nearly the same. For H% labels, `ques` accounted for 40% of the whole data, but the rates for the other markers were also high, especially that of `per` in chiba0232, which was as much as 40%. For HL% labels, `com` occupied 60–70%, and `per` and no marker filled the remaining part. These findings indicate that CSJ's tone labels and CA's intonation markers are not in one-to-one correspondence, and features other than tone labels will be needed for transformation from the CSJ-style transcription into the CA-style transcription.

3.2. Variance in intonation marking between transcribers

Table 1 shows the correspondence between X's and Z's intonation markers in chiba0232 and that between Y's and Z's intonation markers in chiba0432. The agreement between X and Z was 76.0% (κ = .66), which was higher than that between Y and Z (69.9%, κ = .58). Where X and Y placed marker `com`, Z often used an other marker or did not put any marker at all. In addition, for chiba0432, many of the places that were left unmarked in Z's transcription were explicitly marked in Y's transcription. These results indicate that there is considerable variance in intonation marking even between trained transcribers.

3.3. Prediction of intonation markers by the statistical models

Table 2 shows the results for the first-step model predicting the presence of an intonation marker on the basis of the linguistic and acoustic features described in Section 2.3. Because of the variance between the transcribers mentioned in Section 3.2., we used only the data for chiba0232 and chiba0432 transcribed by the same transcriber Z. When one of the two transcriptions was employed as training data, the other served as test data. The accuracies were around 85% for both test data, and the F-measures in predicting the presence of intonation markers were also high (90.6% for chiba0232 and 84.1% for chiba0432, respectively). Figure 4 indicates the relative importance of the predictor variables of this model, which was calculated based on the mean decrease in accuracy. `revLoc` was the most im-

Table 2: Results of predicting the presence of an intonation marker (the first-step model).

Training = chiba0432, Test = chiba0232 (accuracy = 87.4%, $\kappa = .72$)

	Observation	
Prediction	`marked`	`none`
`marked`	**284**	50
`none`	9	**124**

Training = chiba0232, Test = chiba0432 (accuracy = 84.5%, $\kappa = .69$)

	Observation	
Prediction	`marked`	`none`
`marked`	**249**	43
`none`	51	**263**

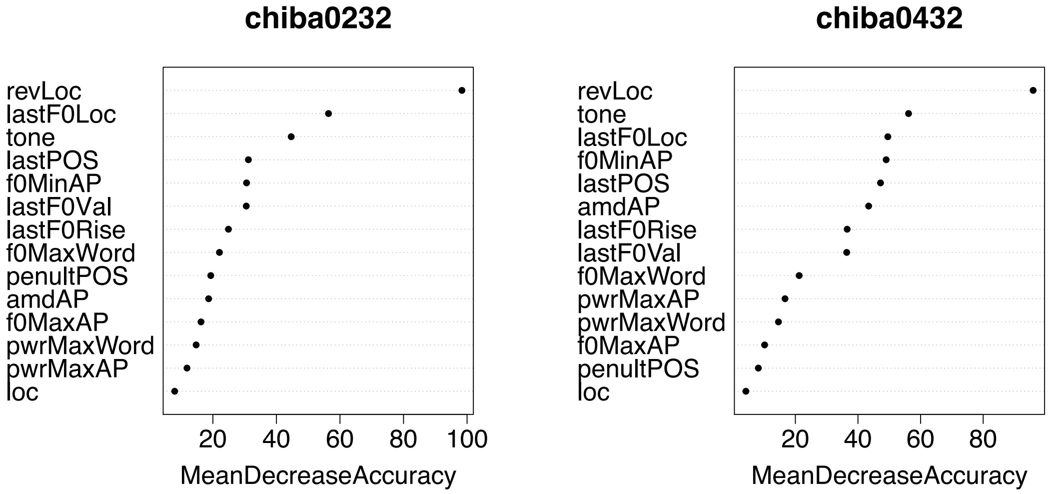

Figure 4: Variable importance for the first-step model.

portant variable for both chiba0232 and chiba0432, and `lastF0Loc` and `tone` were tied for next-most important. While there were some other features that were relatively important in chiba0432, they were less important in chiba0232.

Table 3 shows the results of the second-step model that classifies the types of intonation markers on the basis of the linguistic and acoustic features described in Section 2.3. For this experiment, only those cases that were marked by either of the four intonation markers were used as training and test data, which means that the evaluation is optimistic with the assumption that the first-step model has detected all these cases correctly. The accuracies were relatively high (72.4% for chiba0232 and 72.0% for chiba0432, respectively). There were, however, considerable cases where `ques` was erroneously predicted as `per` in both test data. In addition, `com` was frequently predicted as a wrong marker, `per`, in chiba0432. Figure 5 indicates the relative importance of the predictor variables of this model. The priority of `tone` was clear in both data, but the order of the importance of other features differed much according to the data.

4. Discussion

As a general tendency, transcriber Z less often used intonation markers compared with X and Y. In particular, the proportion of the whole accounted for by `none` (no marker) was 31.4% in Y's transcription of chiba0432 but 50.8% in Z's transcription. There appears to be a difference between their transcription strategies, which might be attributed to the difference between their training environments; only Y has experience in Du Bois' transcription convention (Du Bois et al., 1993), which uses a more phonetic-oriented strategy than the ordinary CA convention. In fact, in his interview, Y stated that he first identified intonational phrases and then put intonation markers at the ends of those phrases. Furthermore, where intonation markers were placed, disagreement between transcribers was also observed. One major difference is that Z used less `com` markers than X or Y did. That is, a remarkable variance between transcribers emerges as to which AP boundaries they regard as bearing continuing intonation.

Even within a single transcriber, the linguistic and acoustic features contributing to prediction of intonation markers differ much between the data sets. One reason for this might be related to the prosodic characteristics of individual speakers; one of the speakers in chiba0432 uses a dialect other than the standard Japanese, and his continuing and rising intonations are different from those pronounced by the other speakers. Another reason is that the H% tone performs a variety of functions other than simple interrogative expression. The CSJ's H% contains emphasis expression, and therefore `per` and `ques` might be classified in the tone H%. In contrast, in spite of H%, `ques` was sometimes predicted as `per` by the other features. It is difficult to correctly predict the CA's `ques` from the linguistic and acoustic features this time, and thus we should consider introducing new features or differenct models.

Table 3: Results of classifying the types of intonation markers (the second-step model).

Training = chiba0432, Test = chiba0232
(accuracy = 72.4%, $\kappa = .42$)

	Observation			
Prediction	per	ques	com	ub
per	**167**	58	3	2
ques	3	**20**	0	0
com	12	2	**25**	1
ub	0	0	0	**0**

Training = chiba0232, Test = chiba0432
(accuracy = 72.0%, $\kappa = .48$)

	Observation			
Prediction	per	ques	com	ub
per	**153**	13	40	5
ques	15	**17**	8	0
com	3	0	**46**	0
ub	0	0	0	**0**

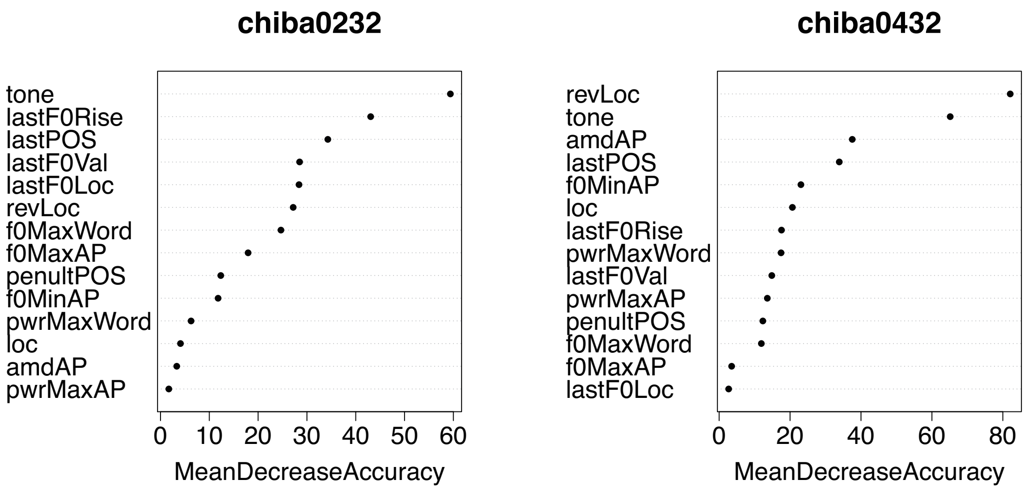

Figure 5: Variable importance for the second-step model.

In sum, we found considerable variance between transcribers as well as between data in the CA-style transcription, which is a significant hurdle in automatic transformation from the CSJ-style transcription to the CA-style transcription. In future work, we will incorporate individual transcription strategy into models and improve the accuracy of the prediction.

5. Acknowledgements

This work was supported by Grant-in-Aid for Collaborative Research Project of NINJAL "Sharing of conversation corpora that cover diverse styles and settings" and JSPS Grant-in-Aid for Scientific Research Number 25370505.

6. References

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Den, Y. and Enomoto, M. (2007). A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation. In Nishida, T., editor, *Conversational informatics: An engineering approach*, pages 307–330. John Wiley & Sons.

Den, Y., Koiso, H., Maruyama, T., Maekawa, K., Takanashi, K., Enomoto, M., and Yoshida, N. (2010). Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC2010)*, pages 2103–2110, Valletta, Malta.

Du Bois, J. W., Schuetze-Coburn, S., Cumming, S., and Paolino, D. (1993). Outline of discourse transcription. In Edwards, J. A. and Lampert, M. D., editors, *Talking data: Transcription and coding in discourse research*, pages 45–89. Lawrence Erlbaum, Hillsdale, NJ.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In Lerner, G., editor, *Conversation analysis: Studies from the first generation*, pages 13–31. Amsterdam: John Benjamins.

Koiso, H., Nishikawa, K., and Mabuchi, Y. (2006). Transcription text (in Japanese). In National Institute for Japanese Language and Linguistics, editor, *Construction of the Corpus of Spontaneous Japanese*, pages 23–132.

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

Maekawa, K. (2003). *Corpus of Spontaneous Japanese*: Its design and evaluation. In *Proceedings of the ISCA and IEEE Workshop on Spontaneous speech processing and recognition (SSPR-2003)*, pages 7–12, Tokyo.