Analysis and Prediction of Morphological Patterns of Backchannels for Attentive Listening Agents

Takashi Yamaguchi, Koji Inoue, Koichiro Yoshino, Katsuya Takanashi,
Nigel G. Ward, Tatsuya Kawahara

Abstract Backchannels play an important role in smooth dialogue, especially attentive listening. In this paper, we analyze the morphological patterns (category) of backchannels, and how these relate to linguistic features of the preceding utterance. In particular we consider the type of the previous utterance-end boundary, the linguistic complexity of the previous utterance, and other features. Based on this analysis, we conduct machine learning to create a model to predict a backchannel's morphological pattern from the preceding context. This model outperforms a baseline: its output better matches the actual backchannels made by human counselors, and human listeners rate its output as more natural.

1 Introduction

There is a growing interest in not only task-oriented dialogue systems but also chatting-style dialogue systems in recent years. One of the applications is conversational agents which conduct attentive listening [1, 2]. In attentive listening such as counseling, it is necessary to give an occasional feedback to the speaker. In spoken dialogue, verbal backchannels such as "*okay*" and "*right*" convey feedback. Without feedback, the speaker would be anxious whether the communication is well maintained. Backchannels suggest that the listener is listening, understanding, and agreeing to the speaker. Backchannels can also express listener's feeling such as interest, surprise and sympathy. These roles are related to the phonetic, morphological

Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, Tatsuya Kawahara
Graduate School of Informatics, Kyoto University, Kyoto, Japan

Koichiro Yoshino
Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan

Nigel G. Ward
Department of Computer Science, University of Texas at El Paso, Texas, USA
Academic Center for Computing and Media Studies, Kyoto University, Kyoto, Japan

and prosodic patterns of the backchannels [3, 4].

Another function of backchannels is to make a comfortable rhythm in the entire conversation. By making synchrony, the speaker feels comfortable in keeping the conversation. In order to make communication smooth, it is very important to generate appropriate backchannels adaptive to the speaker's utterances. On the other hand, the current spoken dialogue systems usually generate the same or limited patterns of backchannels such as "*uh-huh*".

The goal of this study is to generate backchannels depending on the context of the dialogue. This issue should be addressed by both morphological and prosodic aspects, and the former is focused in this paper. We investigate the relationship between the morphological patterns of backchannels and the features of the preceding utterance. First, we focus on the relationship between the boundary type and the morphological patterns of backchannels. We also analyze the relationship between the syntactic complexity and the morphological patterns of backchannels. Then, we investigate whether we can predict the morphological pattern of a backchannel from the preceding utterance based on findings. Finally, we evaluate our prediction model using human subjects.

2 Analysis and Generation of Backchannels

A backchannel is a short response generated by the listener during a dialogue, usually at the end of utterances, without taking a turn; instead, backchannels indicate that the listener does not take a turn. Previous research has focused on three aspects: timing, prosody and lexical form, as these are important to generate natural backchannels in spoken dialogue.

2.1 Timing

Timing of backchannels is usually constrained at the end of the current speaker's utterances, which are defined by a short pause. There are many studies that investigate the timing of backchannels. Ward et al. [5] noted that a low pitch region is one of the major prosodic cues preceding backchannels. Koiso et al. [6] and Noguchi et al. [7] introduced a decision tree based on prosodic and morphological patterns. There are several works that refine prosodic features [8, 9, 10]. There are also several studies which actually implemented a dialogue system to generate backchannels using a decision tree [11, 12, 13]. Recently, an efficient learning mechanism using the wisdom of crowds has been introduced [14].

2.2 Prosody

The prosody of backchannels expresses the listener's feeling such as interest, surprising and sympathy [4, 15]. Heldner et al. [16] showed that the pitch of backchannels is more similar to the immediately preceding utterance compared with normal turn-taking. Kawahara et al. [17] investigated prosodic correlations between the backchannels and the preceding utterances, and showed that the prosody of backchannels synchronizes the prosody of the preceding utterance. Moreover, they tried to predict the prosodic parameters of backchannels based on the property.

2.3 Morphological Patterns (Lexical Entry)

The lexicon of backchannels is language-dependent. In general, backchannels can be divided into many categories. One useful division is between content-rich backchannels, which react strongly to the new information, such as "*wow*" in English and "*hee*" in Japanese: we will call these "reactive" tokens (not to be confused with the general use of "reactive tokens" to mean backchannels in the conversation analysis tradition). More common are backchannels which just show a continued listening posture called "continuer" [18]; these are generally shorter and simpler forms, similar to acknowledge tokens, such as "*yeah*" and "*uh-huh*" in English and "*hai*" and "*un*" in Japanese. Within this latter category, reduplication is commonly observed across languages. English for example has "*uh*", "*uh-huh*", and "*uh-huh-uh-huh*", and Japanese has "*un*", "*un-un*", and "*un-un-un*".

While there are a number of studies on timing and prosody of backchannels, this work focuses instead on analysis of morphological patterns of backchannels. This would be useful for generating the most appropriate pattern of backchannels according to the dialogue context.

3 Analysis on Counseling Dialogue

In this work, we use a corpus of counseling dialogue [17]. These are not real counseling, in that the subjects were asked to come to the session for dialogue data collection, not for counseling. But they were asked to talk about their real personal troubles, for example, human relationship and career path, to a counselor. The subjects are eight college students. We had two counselors, and each took part in four sessions. All participants are Japanese native. The counseling session lasted around 20-30 minutes. The annotation of backchannels was done by an expert according to a guideline.

In this work, we focus on the distributional differences, first between the reactive tokens and the continuers, and second between base forms and reduplicated forms

of the continuers. As reactive tokens we consider the forms "*haa*", "*aa*" and "*hee*". For tokens of "*un*" and "*hun*" and their reduplicated forms of continuers, we also distinguish between simple doubling and repeats of 3 or more. Other backchannels such as "*uuun*" were excluded from the analysis because they are minor and do not fall into these categories. In summary, we have four categories of backchannels: un, unx2, unx3 and reactive.

In English, reduplication of backchannels has been associated with indicating a lack of anything to add, and thus demonstrating the intention to not take the floor, even if an opportunity is present [19]. Ward [20] suggested the same was true for Japanese. Reduplication in Japanese backchannels was also examined by Katagiri et al. [21], who characterized the reduplicated forms as indicating "emphasis". Experimentally they found a slight tendency for the reduplicated forms to be perceived as more lively and involved. In this work, we do not attempt to identify any specific meaning for the reduplicated forms, instead we examine how the choice of form is directly affected by the local context.

Many backchannels overlap continuing talk by the other speaker or occur without any significant intervening pause. In such contexts the simplest forms, "*un*" and "*hun*", are overwhelmingly used. On the other hand, after clause boundaries, a full variety of forms is employed. We accordingly focus on backchannels appearing after clause boundaries, which are defined by some linguistic analysis, mostly by the end-of-clause patterns in Japanese.

3.1 Relationship between Boundary Type and Morphological Patterns of Backchannels

First, we investigate the relationship between the category of backchannels and the boundary type of the preceding utterance. The boundaries are annotated as one of three types according to the guideline of the Corpus of Spontaneous Japanese (CSJ) [22]: absolute boundary, strong boundary and weak boundary. The absolute boundary is the end of the sentence. The strong boundary is not the end of the sentence, but the preceding clause and the following clause do not have a depending relations e.g. they are parallel clauses. The weak boundary is a boundary between clauses which have a dependency relation. We counted the frequency of the backchannel category for each boundary type in Table 1. The result indicates that reactive and unx3 appear more frequently in the absolute boundary than other categories of un and unx2. On the other hand, un and unx2 are frequently used in strong and weak boundaries, and there is no clear difference in the occurrence tendency between their boundaries.

Table 1 Counts of backchannel categories for different types of clause boundary.

	un	un x2	un x3	reactive	total
weak boundary	**141** (28.5%) (68.4%)	**81** (16.4%) (62.3%)	**161** (32.5%) (48.8%)	**112** (22.6%) (38.2%)	495
strong boundary	**48** (24.9%) (23.3%)	**33** (17.1%) (25.4%)	**69** (35.8%) (20.9%)	**43** (22.1%) (14.7%)	193
absolute boundary	**17** (6.3%) (8.3%)	**16** (5.9%) (12.3%)	**100** (36.9%) (30.3%)	**138** (50.9%) (47.1%)	271
total	206	130	330	293	959

The percentage in the upper row is computed for backchannel categories and the percentage in the lower row is computed for clause boundary types.

Table 2 Statistics of syntactic features in the preceding clause/sentence unit averages.

	un	un x2	un x3	reactive
number. of phrases	**4.73**	**5.52**	5.42	5.15
depth of phrases	**2.18**	**2.57**	2.56	2.54
width of phrases	1.88	2.00	1.89	1.75

3.2 Relationship between Syntactic Complexity and Morphological Patterns of backchannels

We also conduct an analysis of relationships between the linguistic complexity of the preceding utterance and the category of the backchannels. We counted the number of phrase units (bunsetsu) in the preceding utterance, and the depth and width of their dependency. A Japanese dependency parser KNP [1] is used for this analysis. The width means the number of phrases directly depending on the last word of the preceding utterance (usually verb or event noun in Japanese), and the depth means the maximum number of edges in the depending tree. We calculated these measures as indications of the complexity of the preceding utterance. An examples is presented in Fig. 1. Statistics aggregated by the backchannel categories are shown in Table 2. As a result of the pair-wise t-test ($p < 0.05$) between the categories, there is a statistically significant difference between un and unx2, shown in bold fonts, in the average number and the depth of the phrase dependency. The result suggests that we can predict the category of un or unx2 by using the complexity of the preceding utterance. Note that this distinction was not possible by considering only boundary strength as discussed in the analysis of Section 3.1.

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

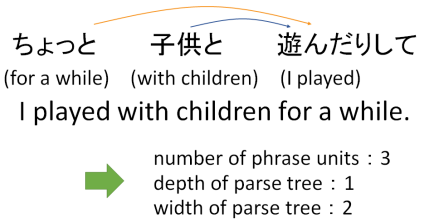

Fig. 1 Examples of a sentence with parsed result

4 Prediction of Morphological Patterns of Backchannels

Next, we conduct prediction of the morphological category based on features of the preceding utterance. The prediction selects from the four categories: un, unx2, unx3 and reactive.

4.1 Annotation of Acceptable Backchannels

In the corpus of counseling dialogue, usually only one token of a backchannel appears in each context, namely the one produced by the counselor. In many cases, however, other categories of backchannels are also acceptable and might have been used by other persons. For example, when a person uses a reactive backchannel "*aa*" to express strong agreement, "*un un un*" is also acceptable to express the similar feeling. To deal with this problem, we additionally annotated other acceptable backchannel categories for every backchannel occurrence of the test set. Three annotators were engaged, and each annotator was given a preceding utterance and evaluated the acceptability of each backchannel category by using a 3-level scale: 1) acceptable, 2) difficult to decide, 3) not acceptable.
The experimental procedure is as follows.

1. An annotator listens to the entire counseling session from the beginning.
2. The system prompts annotation when the playing reaches the timing of a backchannels that actually appeared in the session. (the actual backchannels is not played at this moment.)
3. The annotator rates the acceptability of the four categories of backchannels.

The agreement among the three annotators is very low, suggesting the use of backchannels category is very person-dependent. Therefore, we adopt a "safe" criterion. When a backchannel category is annotated as acceptable by all three annotators, we regard it as a "correct" label for the reference in evaluating prediction of the backchannel category. The average number of acceptable category per token is 2.3. In this process, un and unx2 were added at most of the spots since these are acceptable in most cases.

Table 3 Features used for prediction

Type of feature	Example
boundary strength type	weak boundary
last word	*de*
part of speech of last word	auxiliary verb / continuative form
boundary syntactic-type label	*de-setsu*
number of phrases	3
depth of phrase	2
width of phrase	1
boundary strength type before previous backchannel	weak boundary
morphological pattern of previous backchannel	*un*

Table 4 Training data and test data for prediction [2]

Train data	*un*	*un*x2	*un*x3	reactive	Test data	*un*	*un*x2	*un*x3	reactive
w/o Session 1	135	89	234	183	Session 1	14	31	30	18
w/o Session 2	127	121	252	186	Session 2	19	3	24	21
w/o Session 3	150	97	226	176	Session 3	2	27	36	20
w/o Session 4	122	125	253	188	Session 4	15	3	10	15

4.2 Experimental Settings

In this experiment, we classify the backchannel category using features of the preceding utterance. We selected features based on the analysis of Section 3, which are described in Table 3. We also used part of speech, the last word of the preceding utterance and the boundary label which defines the boundary type. These are extracted from the manual transcripts. In addition to these, the morphological pattern of the previous backchannel and its clause boundary type were used. We used linear support vector machines (SVM) for machine learning [23].

We used eight sessions of the counseling dialogue as training data and out of the these four sessions with the best recording conditions as testing data, which were annotated in the previous subsection. Since these four sessions are a subset of the training data, classification experiments were carried out using a leave-one-session-out manner. When one session is used for testing, it is held out from the training data. The details of the data are shown in Table 4.

For reference, we assume a system which always output unx3, which is the most frequent, as a baseline. The prediction performance is evaluated with precision, recall and F-measure, which is the harmonic mean of the precision and recall.

[2] The counts in the test data do not include tokens added in Section 4.1.

Table 5 Prediction performance of backchannel category.

	Prediction model	Baseline
Session 1	**0.591 (55/93)**	0.376 (35/93)
Session 2	**0.582 (39/67)**	0.448 (30/67)
Session 3	**0.647 (55/85)**	0.529 (45/85)
Session 4	**0.651 (28/43)**	0.233 (10/43)
Average	**0.615**	0.417

Table 6 Prediction performance for each backchannel category.

backchannel	Precision	Recall	F-measure
un	0.917 (44/48)	0.420 (21/50)	0.576
un x2	0.879 (29/33)	0.375 (24/64)	0.526
un x3	0.444 (48/108)	0.780 (78/100)	0.566
reactive	0.566 (56/99)	0.730 (54/74)	0.637
All	0.615 (177/288)	0.615 (177/288)	**0.615**

4.3 Evaluation of Prediction Performance

The performance of prediction is shown in Table 5. The prediction model by machine learning using the features of the preceding utterance outperformed the baseline for all sessions. The performance for each backchannel category is listed in Table 6. The proposed prediction model also performed better than the baseline in all categories. These results show that the proposed model realizes appropriate selection of backchannels, which fit the context. The usefulness of the analysis results of Section 3 is also suggested by this result. However, precision of reactive is low, suggesting it was predicted in inappropriate spots. As these are rich in specific meaning, improper generation of these could have a large effect on perceptions. The problem needs to be solved before implementing a dialogue system.

5 Evaluation of Proposed Model by Human Subjects

5.1 Experimental Settings

Next, we evaluate our prediction model by human subjects. The subjects in this experiment are nine persons (five male and four female). We asked the subjects to listen to dialogue segments and make subjective evaluation on the backchannels. For comparison, we inserted backchannels predicted by following three methods into the clause boundary where the actual backchannels are observed in the corpus. We adopted unx3 as the baseline, which was used in the previous section.

1. Baseline：always generate unx3

2. Prediction : generate a pattern predicted by the proposed model using a fixed recorded sample for each token. For reactive tokens, always generate "*haa*".
3. Counselor : use the backchannel pattern actually chosen by the counselor and replace it with a fixed recorded sample of that token.

The patterns of the backchannels used in this experiment include "*un*", "*un un*", "*un un un*", "*haa*", "*aa*". These are most acceptable tokens in each category, as shown in Section 4.1. We extracted eight segments from four sessions used in the previous section where each segment lasted around two minutes. In these segments, the content of the story are coherent and the counselor mostly utters backchannels only. The occurrence frequency of the backchannels is about ten in each segment. Since these segments begin from the middle of the dialogue, we explained an outline of each session to the subjects before they started the experiment.

As there were 24 stimuli of the 8 segments and the 3 backchannel generating methods, each stimulus was assigned to 3 different subjects. When we assigned the sample segments to each subject, we minimized the duplication of combination of same subjects among different slots. In addition, the subjects listened to the segments in different orders.

We prepared evaluation items by considering the functions of backchannels defined in [24]. The subjects conducted five-grade evaluations (-2: Strongly Disagree, -1: Disagree, 0: Don't know, 1: Agree, 2: Strongly Agree).

5.2 Subjective Evaluation Result

The evaluation results are shown in Table 7. The figures present the average values from all subjects. Evaluation scores were, from high to low, in the order of counselor, prediction, baseline, in all questions. As a result of the pair-wise t-test ($p < 0.05$), there are significant differences between the baseline and the prediction model, shown in bold fonts, in Q1, Q4, Q5, Q6. However, there are not significant differences between the prediction model and the counselor in all questions

We observe very low evaluation of the baseline in Q1. When the morphological pattern of backchannels is always same, the listener has an unnatural impression. Also in Q2, the variety of morphological patterns of the backchannels gives an impression that the dialogue has a good tempo.

In item Q3 related to the function of backchannels, there are not significant differences between the baseline and the prediction model. The result suggests that even the simple repetition of backchannels can give an impression that you are willing to listen to the talk. In Q4 and Q5, however, the evaluation scores are lower for all, even the human-chosen forms. Nevertheless, the prediction model is significantly above the baseline.

Finally, Q6 is a comprehensive evaluation on the implementation of an attentive listening system. The result is similar to the trend of the evaluation scores in Q4 and Q5.

Table 7 Evaluation results of backchannel generation methods

Question	Baseline	Prediction	Counselor
Q1: Did you feel backchannels is natural in a whole conversation ?	**-0.583**	**0.208**	0.458
Q2: Did you feel this conversation has a good tempo ?	0.000	0.417	0.792
Q3: Did you feel that this counselor listens to the talk ?	0.042	0.417	0.542
Q4: Did you feel that this counselor understands the talk ?	**-0.625**	**0.083**	0.208
Q5: Did you feel that this counselor sympathizes with the talk ?	**-0.750**	**-0.083**	0.375
Q6: Did you want to talk with this counselor ?	**-0.667**	**-0.083**	0.458

In general, a listener cannot understand the talk without listening and cannot sympathize without understanding [25]. Hence, we believe that there is a degree of depth in understanding of the listener and it is necessary that the listener expresses the depth of understanding by the strength of the reaction. From this point, it is not possible to express strong reaction such as deep understanding and sympathy by continuing the same morphological pattern of backchannels, and our model indeed performs better than the baseline.

On the other hand, the prediction model has lower evaluation than the counselor. We reason this is because reactive tokens having a function of understanding and sympathy sometimes appeared at inappropriate spots by the prediction model. This result corresponds with the fact that precision of reactive tokens was low in the previous section of the prediction experiment.

6 Conclusions

We have investigated effective generation of backchannels depending on the context of dialogue. First, we analyzed the relationship with the boundary type of the preceding utterance. Second, we analyzed the relationship with the syntactic complexity of the preceding utterance. Based on these analyses, we made prediction of the morphological patterns of backchannels from the preceding utterance. Our prediction model showed better accuracy than the baseline method that always outputs the most frequent pattern, in two ways: correspondence to labels and human judgments. Both evaluation results show that our method obtained a significantly higher rating than the baseline method.

For future work, we will develop a system which generates backchannels depending on the context of the dialogue based on these findings.

Acknowledgment

This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project.

References

1. Y. Kobayashi, D. Yamamoto, T. Koga, S. Yokoyama, and M. Doi. Design targeting voice interface robot capable of active listening. In *Proc. HRI*, pp. 161–162, 2010.
2. D. DeVault, R. Artstein, G. Benn, T. Dey, G. E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proc. AAMS*, pp. 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems, 2014.
3. N. Ward. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody*, pp. 325–328, 2004.
4. T. Kawahara, Z. Q. Chang, and K. Takanashi. Analysis on prosodic features of Japanese reactive tokens in poster conversations. *Speech Prosody*, pp. 1–4, 2010.
5. N. Ward and W. Tsukahara. Prosodic features which cue backchannel responses in English and Japanese. *J. Pragmatics*, Vol. 32, No. 8, pp. 1177–1207, 2000.
6. H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech*, Vol. 41, No. 3, pp. 295–321, 1998.
7. H. Noguchi and Y. Den. Prosody-based detection of the context of backchannel responses. In *Proc. ICSLP*, Vol. 2, pp. 8570–8573, 1998.
8. T. Solorio, O. Fuentes, N. Ward, and Y. A. Bayyari. Prosodic Feature Generation for Back-Channel Prediction. In *Proc. Interspeech*, pp. 2398–2401, 2006.
9. K. P. Truong, R. Poppe, and D. Heylen. A Rule-Based Backchannel Prediction Model Using Pitch and Pause Information. In *Proc. Interspeech*, pp. 3058–3061, 2010.
10. A. Gravano and J. Hirschberg. Backchannel-Inviting Cues in Task-Oriented Dialogue. In *Proc. Interspeech*, pp. 1019–1022, 2009.
11. R. Nishimura, N. Kitaoka, and S. Nakagawa. A spoken dialog system for chat-like conversations considering response timing. In *Text, Speech and Dialogue*, pp. 599–606. Springer, 2007.
12. N. Kitaoka, M. Takeuchi, R. Nishimura, and S. Nakagawa. Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. *J. Japanese Society for Artificial Intelligence*, Vol. 20, pp. 220–228, 2005.
13. S. Fujie, K. Fukushima, and T. Kobayashi. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. In *Proc. INTERSPEECH*, pp. 889–892, 2005.
14. D. Ozkan and L.-P. Morency. Modeling wisdom of crowds using latent mixture of discriminative experts. In *Proc. ACL/HLT*, 2011.
15. T. Kawahara, S. Hayashi, and K. Takanashi. Estimation of interest and comprehension level of audience through multi-modal behaviors in poster conversations. In *Proc. INTERSPEECH*, pp. 1882–1885, 2013.
16. M. Heldner, J. Edlund, and J. Hirschberg. Pitch Similarity in the Vicinity of Backchannels. In *Proc. Interspeech*, pp. 3054–3057, 2010.
17. T. Kawahara, M. Uesato, K. Yoshino, and K. Takanashi. Toward adaptive generation of backchannels for attentive listening agents. In *Proc. International Workshop on Spoken Dialog Systems*, 2015.
18. E. Schegloff. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. *Analyzing discourse: Text and talk*, pp. 71–93, 1982.
19. R. Gardner. The Conversation Object Mm: A weak and variable acknowledging token. *Research on Language and Social Interaction*, Vol. 30, No. 2, pp. 131–156, 1997.
20. N. Ward. The relationship between sound and meaning in Japanese back-channel grunts. *Proceedings of the 4th Annual Meeting of the (Japanese) Association for Natural Language Processing*, pp. 464–467, 1998.
21. Y. Katagiri, M. Sugito, and Y. Nagano-Madsen. The Forms and Prosodic Characteristics of Backchannels in Tokyo and Osaka Japanese. In *Proc. ICPhS*, pp. 2411–2414, 1999.

22. K. Maekawa. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 7–12, 2003.
23. R. E. Fan, K. E. Chang, C. H. Hsieh, X. R. Wang, and C. J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874, 2008.
24. J. Horiguchi. Linguistic behavior of the listener in communication (in Japanese). *J. Teaching Japanese as a Foreign Language*, No. 64, pp. p13–26, 1988.
25. J. Allwood, J. Nivre, and E. Ahlsén. On the semantics and pragmatics of linguistic feedback. *The Journal of semantics*, Vol. 9, No. 1, pp. 1–26, 1992.