Where in Dialog Space does *Uh-huh* Occur?

Nigel G. Ward, David G. Novick, Alejandro Vega

Department of Computer Science, University of Texas at El Paso, El Paso, Texas, United States

nigelward@acm.org, avega5@miners.utep.edu, novick@utep.edu

Abstract

In what dialog situations and contexts do backchannels commonly occur? This paper examines this question using a newly developed notion of dialog space, defined by orthogonal, prosody-derived dimensions. Taking 3363 instances of *uh-huh*, found in the Switchboard corpus, we examine where in this space they tend to occur. While the results largely agree with previous descriptions and observations, we find several novel aspects, relating to rhythm, polarity, and the details of the low-pitch cue.

Index Terms: backchannels, feedback, prosody, context, principal component analysis, dimensions, dialog activities

1. The Contexts of Backchannels

Among the interactive phenomena of dialog, backchanneling is one of the most prototypical and among the most studied. A key question of interest is when backchannels occur. Some aspects of this question have been intensively investigated, for example regarding the prosodic contexts that cue backchannels and similar feedback [1, 2].

While we know some things about the micro-contexts of backchannels in certain situations, we lack a good understanding of the more general dialog situations in which backchannels occur. Indeed, descriptions at this level tend to come not from empirical study but from definitions, theoretical frameworks, qualitative studies, and impressionistic observations. Aspects of situations often thought to be relevant to backchanneling include having only one person holding the floor, giving a narrative or explanation, having one person being in a listening role, having that person supportive and maybe even agreeing, and being at points of new information or where grounding needs to be done [3, 4]. However such listings of factors have lacked empirical verification, may include properties that are rarely important, and may omit properties that are vital in practice.

This is a problem for efforts to build and deploy responsive systems. It is possible to backchannel naturally and effectively in dialogs with naive humans [5, 6], but so far only when the user's role is tightly constrained, for example to retell a story or solve a simple problem. To make backchanneling behavior (and ultimately other types of rapid response behaviors) more robust and useful in freer dialog settings, we need a better understanding of the dialog contexts and activities in which they occur.

Thus we here undertake an empirical, statistical exploration of the dialog contexts of occurrence of backchannels.

2. Dialog Dimensions

To understand the typical dialog situations where backchannels occur, we need to start with a way to describe dialog situations. While there are many taxonomic systems to choose from, here we use a new, empirical method [7]. Reasoning that the local prosody is a good indicator of dialog activities and states, we started with 76 local prosodic features, consisting of pitch height, pitch range, speaking rate, and volume, computed over different regions of a 6 second window, and computed for both participants in the dialog. We computed these features every 10 milliseconds throughout the corpus. We then applied Principal Component Analysis to these values. This gave a list of 76 dimensions, ordered by how much of the variation in the prosodic features they explain.

Upon examination [7, 8], most of the top dimensions turned out to align with aspects of dialog. These aspects were diverse, including dialog situations, transient dialog states, cooperative dialog acts, simpler dialog actions, apparent mental states, and some prosodic behaviors.

Since these are truly dimensions, they are continuously valued. Thus, a given moment in dialog might have a value –0.74 on dimension 1, +.03 on dimension 2, and so on. Any specific instance of a backchannel occurs at a point in this 76-dimensional dialog space, and thus we can gather statistics on where backchannels tend to occur.

3. *Uh-Huh* and the Dimensions

To determine the typical dialog contexts of backchannels, we examined the patterns of occurrence of *uh-huh* in the Switchboard corpus. We chose Switchboard because it comprises unstructured dialogs and includes a wide variety of dialog activities. We chose *uh-huh* as a proxy for backchannels because *uh-huh* is almost always a backchannel and is one of the most common typical backchannel forms (along with *uh, yeah, [laughter], oh*,

and *um-hum*). We gathered statistics over a 600K word subset of Switchboard, which includes 3363 instances of *uh-huh*. These tokens are not limited to phonetically precise *uh-huh* tokens, as the labelers guide enjoins transcribers to "use "uh-huh" or "um-hum" (yes) ... for anything remotely resembling these sounds of assent" [9] (although in practice the degree of yes-ness did not seem to much affect what ended up in the transcripts).

If backchannels were mostly about reacting to content or mostly about reacting to a few specific cues, we would expect them to be distributed evenly across most of the dimensions. However in fact many of the distributions were strongly asymmetric: for twelve of the dimensions 75% or more of the occurrences of *uh-huh* were on one side or the other (positive or negative) of the dimension, as seen in the left columns of Table 1, and all of these asymmetries are significant (by the chi-square test, $p < .0001$). Thus backchannels do relate to multiple dimensions of dialog.

4. Interpretations

While the statistical information in Table 1 is adequate as a practical, operational answer to the question of where backchannels occur, we wanted to go further, to develop a real understanding of the reasons for and significance of the associations with these regions of dialog space.

For this, our initial exploration, we examined each dimension individually. We generally started with a previous description of the dimension [7, 8] and tried to understand how *uh-huh* was similar to the other dialog activities that were common in those particular contexts. To do this we listened to examples of speech in the vicinity of *uh-huh* as it appeared in those contexts and to examples of other words in similar contexts. While exploratory inductive studies of this sort risk merely confirming the observers' prejudices, that was not the case here. Indeed, we were repeatedly surprised to see connections to constructs that had we had not previously thought relevant, notably transition relevant places and dialog rhythms.

4.1. The 12 Most-Related Dimensions

From the twelve dimensions where the distribution of *uh-huh* was most skewed, we infer connections to:

Turn grabbing

91% of the occurrences of *uh-huh* were in contexts that were low on dimension 5. In other situations low on this dimension the speaker is starting a turn. (Situations high on this dimension were mostly turn yields.) This implies that *uh-huh* can sometimes take the turn and also may function to decline to take a turn at a point when that opportunity was available [4].

Pushing for a new perspective

89% of the cases of *uh-huh* occurred in contexts on the lower half of dimension 17. Other typical dialog actions low on dimension 17 were short questions or other swift bids to slightly change the topic while the interlocutor is monopolizing the floor. (In situations high on this dimension the speaker and/or interlocutor were generally engaged in elaborating a feeling or mood.) This suggests that *uh-huh* can be a way to move the conversation forward, a facet which the common term "continuer" highlights.

Quick thinking

89% were on the high side of dimension 11. Typical utterances high on dimension 11 were very swift echos and confirmations. (Situations on the low side were typically low in confidence and/or content, for example when ending an utterance with *I guess*.) This suggests that *uh-huh* can indicate attention and quick understanding.

Expressing sympathy

86% were high on dimension 18. In other typical dialog situations high on this dimension the speaker is expressing pity or sympathy for someone in a bad situation the interlocutor has just described. (Situations low on this dimension were frequently descriptions of people or happenings meriting sympathy, and thus soliciting an expression of sympathy.) This suggests that *uh-huh* can convey sympathy.

Expressing empathy

86% were high on dimension 6; this region typically included expressions of empathy. To clarify, this dimension differs from the previous one in that it includes positive emotions and evaluations. In terms of prosodic contexts, empathy typically respond to a phrase or word produced in high pitch by the interlocutor (*Arizona's beautiful*), whereas sympathetic responses usually respond to phrases produced in low volume and reduced pitch range. (Situations low on this dimension were typically emotional expressions and evaluations, which often invited an expression of empathy.) Thus *uh-huh* patterns with other expressions of empathy.

Other speaker talking

85% were high on dimension 1. In other situations high on this dimension the interlocutor was talking almost constantly while the speaker of interest was mostly quiet. (The low end of this dimension was the exact opposite.) Interestingly, here the simple percentage does not tell the whole story: in fact 67% of the *uh-huh*s were in the 3rd quartile on this dimension. This indicates, unsurprisingly, that *uh-huh* occurs when it is mostly the interlocutor who is speaking, but not in an extreme monolog context. This facet is naturally the one which the term "backchannel" highlights.

Rambling

82% of the cases of *uh-huh* occurred in contexts that were on the lower half of dimension 14. Other typical dialog situations low on this dimension were where the speaker has low interest in what he himself is saying, but

seems to feel the need to say something anyway. (On the high side, the speaker was usually speaking clearly, even emphatically, in a bright tone.) This suggests that *uh-huh* can convey low interest and a lack of anything specific to say, a facet which the common term "minimal vocalization" highlights.

Signaling an upcoming point of interest

79% were high on dimension 26. At points with high values on this dimension, the speaker often seems to be signaling that the dialog is about to take off in some way. Prosodically, this is characterized by a moderately high volume for a few seconds that then turns low and is accompanied by a slower speaking rate and a region of low pitch for a hundred milliseconds or so; after this comes the point of interest and then in the near future typically both speakers have some speaking role, both with higher than average pitch height. (At points with low values on this dimension, the speaker is typically involved in a narrative and speaking with low volume, and appears to be downplaying the importance of what he's saying, for example in situations where he needs indicate that what he's saying is just background to an upcoming main point.) Thus *uh-huh* can be cued by a prosodic context including a region of low pitch, which elaborates a well-known result [10].

Deploring something

78% were high on dimension 37. At other points with high value on this dimension, the speaker is often describing something deplorable, as in *if the legislature has their way about it they're going to ráise the tuítion and dóuble* and in *the stráw that bróke the cámel's back*, with something of a sing-song unstressed-stressed alternation. (Times with low values on this dimension often fall near the point where the speaker starts to reveal that a situation also has a silver lining.) This suggests that an *uh-huh* can serve to share a complaint.

Not delivering confidently

76% were low on dimension 72. At other points low on this dimension the speaker's delivery was often weak, in the extreme including false starts or disfluencies, and the interaction between the speakers, if any, was awkward. (At points high on this dimension the speaker had established something of a rhythm of speaking although, unlike the previous dimension, typically with several unstressed syllables between each stressed syllable. If the listener was saying anything at all, his words tended to fit in smoothly where the speaker would have put unstressed syllables. Pragmatically, this seems common in cases where the speaker really knows what he wants to say.) Thus *uh-huh* patterns with speaking without full confidence and a clear delivery.

Agreeing and preparing to move on

76% were high on dimension 24. In other situations high on this dimension, the speaker was expressing agreement with or sharing the other's thought or feeling, preparatory to moving the focus to a new aspect of the topic. (In low situations on this dimension both speakers were focusing for some time on the same shared referent.) Thus *uh-huh* patterns with agreeing, closing out, and bidding to move on.

Low focus

75% were low on dimension 29. In general this was seen in contexts where there was an unstressed or somehow deemphasized word. While we found no consistent pragmatic or dialog function for these, sometimes they co-occurred with taking a personal stance. (On the high side of this dimension there was a stressed word in the context, and this often occurred where the focus was on establishing the facts, and where the speaker had knowledge that the interlocutor clearly lacked.) Thus *uh-huh* patterns with a lack of stress, a relative lack of knowledge of the topic, and with taking a stance that is personal, rather than fact-oriented.

4.2. Other Dimensions

Table 2 lists others among the top two dozen dimensions for which the distribution of *uh-huh* is strongly asymmetric; discussion of each dimension appears elsewhere [7, 8]. While some are not surprising (the correlations with dimensions 8, 12, and 19), dimension 10 is more interesting: the propensity of low values indicates that *uh-huh* patterns with thinking of something to say next, rather than being disengaged from the dialog. This is perhaps why *uh-huh* can work for feigning attention, as Yngve humorously observed [3]. Dimension 13 suggests that backchannels are similar in some ways to the beginnings of contrasts, and dimension 21 to actions done to mitigate potential face threats. These aspects should be looked at more closely.

4.3. Discussion

First, we note that our new analysis method, indirect thought it is, appears to work: most of the aspects of dialog that *uh-huh* co-occurs with are things that could be expected from one or another of the common descriptions of backchannel. Yet the results were not without novelty, for example in the connections with the dialog aspects involved in dimensions 14, 37, 72, 29, 13, and 21.

Second, we note the wide variety of factors involved in backchannel behavior. Although many studies of these phenomena approach them in ways that limit the findings to just one type of context, or one type of cue, or one type of effect, in fact backchannels are richly multifaceted and multifunctional.

5. Future Work

Tracking the dialog situation using the dimensions identified here may enable future dialog systems to perform backchanneling more robustly, even with uncontrolled

Dimension	Skew	Interpretation (Abbreviated)
PC 5 lo	91%	turn grab (vs. turn yield)
PC 17 lo	89%	pushing for a new perspective (vs. elaborating current feeling)
PC 11 hi	89%	attentive and quick-thinking (vs. low confidence)
PC 18 hi	86%	expressing sympathy (vs. seeking sympathy)
PC 6 hi	86%	expressing empathy (vs. seeking empathy)
PC 1 hi	85%	other speaker talking vs. this speaker talking
PC 14 lo	82%	rambling (vs. placing emphasis)
PC 26 hi	79%	signaling interestingness (vs. downplaying things)
PC 37 hi	78%	deploring something (vs. also planning to talk about the good side)
PC 72 lo	76%	speaker awkward (vs. speaking with a clear delivery)
PC 24 hi	76%	agreeing and preparing to move on (vs. jointly focusing)
PC 29 lo	75%	no recently stressed word (vs. stressed word present)

Table 1: The 12 dimensions with the most asymmetric distributions of *uh-huh*, with interpretations of the *uh-huh*-rich side (vs. the opposite side).

Dimension	Skew	Interpretation (Abbreviated)
PC 8 hi	66%	ending crisply (vs. petering out)
PC 10 lo	73%	engaging in lexical or memory access (vs. disengaged)
PC 12 lo	69%	floor yielding (vs. floor asserting)
PC 13 lo	66%	starting a contrasting statement (vs. reiterating)
PC 19 lo	75%	solicitous (vs. controlling)
PC 21 lo	71%	mitigating a potential face threat (vs. agreeing, with humor)

Table 2: Seven other high-variance dimensions with significant asymmetry.

users, with the likelihood that they the backchannels will occur only in suitable contexts.

While this exploration of dialog space has identified some general areas where *uh-huh*s occur, with respect to each dimension independently, we would like to more tightly characterize these regions. In particular, we would like to explore whether *uh-huh* has distinct subpopulations and, if so, whether these have distinctive phonetic and prosodic properties

More generally, this new analysis method could be used to help discover and characterize the roles and typical contexts of other dialog-relevant markers and behaviors.

6. Acknowledgments

This work was supported in part by NSF Award IIS-0914868. We thank Tatsuya Kawahara for comments.

7. References

[1] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, pp. 70–84, 2010.

[2] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language*, vol. 25, pp. 601–634, 2011.

[3] V. Yngve, "On getting a word in edgewise," in *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pp. 567–577, 1970.

[4] E. A. Schegloff, "Discourse as an interactional achievement: Some uses of "Uh huh" and other things that come between sentences," in *Analyzing Discourse: Text and Talk* (D. Tannen, ed.), pp. 71–93, Georgetown University Press, 1982.

[5] J. Gratch, N. Wang, A. Okhmatovskaia, F. Lamothe, M. Morales, R. van der Werf, and L.-P. Morency, "Can Virtual Humans Be More Engaging Than Real Ones?," *Lecture Notes in Computer Science*, vol. 4552, pp. 286–297, 2007.

[6] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, Gary, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wollmer, "Building autonomous sensitive artificial listeners," *IEEE Transactions on Affective Computing*, to appear.

[7] N. G. Ward and A. Vega, "A bottom-up exploration of the dimensions of dialog state in spoken interaction," in *Sigdial*, 2012.

[8] N. G. Ward and A. Vega, "Towards empirical dialog-state modeling and its use in language modeling," in *Interspeech*, 2012, submitted.

[9] J. Hamaker, Y. Zeng, and J. Picone, "Rules and guidelines for transcription and segmentation of the Switchboard large vocabulary conversational speech recognition corpus, version 7.1," tech. rep., Institute for Signal and Information Processing, Mississippi State University, 1998.

[10] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, pp. 1177–1207, 2000.