A Shallow Model of Backchannel Continuers in Spoken Dialogue

Nicola Cathcart
Canon Research Centre Europe
Bracknell, UK
nicolac@cre.canon.co.uk

Jean Carletta and **Ewan Klein**
School of Informatics
University of Edinburgh
{jeanc,ewan}@inf.ed.ac.uk

Abstract

Spoken dialogue systems would be more acceptable if they were able to produce backchannel continuers such as *mm-hmm* in naturalistic locations during the user's utterances. Using the HCRC Map Task Corpus as our data source, we describe models for predicting these locations using only limited processing and features of the user's speech that are commonly available, and which therefore could be used as a low-cost improvement for current systems. The baseline model inserts continuers after a predetermined number of words. One further model correlates back-channel continuers with pause duration, while a second predicts their occurrence using trigram POS frequencies. Combining these two models gives the best results.

1 Introduction

In a spoken dialogue between people, the participants use simple utterances such as *yeah*, *a totty wee bit aye* and *mm-hmm* to signal that communication is working. Without this feedback, the partner may assume that he has not been understood and reformulate his utterance. Following Yngve (1970), we will use the term *backchannel* for such utterances. Although these can be substantive because they can repeat material from the partner's utterance (Clark and Schaefer, 1991), e.g., *Right, okay, I'm below the flat rocks*, we will adopt (Jurafsky et al., 1998)'s terminology of *continuer*. We will take this to refer to the class of backchannel utterances, with minimal content, used to clearly signal that the speaker should continue with her current turn. (Yankelovich et al., 1995) point out that users of speech interface systems need feedback, too, especially since the system's silence could mean either of two very different things: that it is waiting for user input, in which case the user should speak, or that it is still processing information, in which case the user should not. However, any feedback must come at the right time or else it risks disrupting the speaker and ultimately, delaying task completion (Hirasawa et al., 1999).

Most of our data, including the examples given above, are drawn from the HCRC Map Task Corpus, described in more detail in Section 3. Clearly these dialogues are significantly more complex than the kind of interactions supported by current commercial spoken dialogue systems, where the length of user utterances is severely constrained. What kind of system would involve potentially lengthy user instructions comparable to those found in the Map Task? Lauria et al. (2001), Lemon et al. (2002), and Theobalt et al. (2002) describe work on building spoken dialogue systems for conversing with mobile robots, and this is a setting where complex instructions naturally arise. For example, in one scenario,[1] users attempt to teach routes and route segments to a robot. (1) is a portion of such an instruction.

(1) okay go to the end of the road and turn left and ...erm ...and then carry on down that road and then turn ...take your second left where the trees are on the corner

We describe a shallow model, based on human dialogue data, for predicting where to place backchannel feedback. The model deliberately requires only simple processing on information that spoken dialogue systems already keep as history, and is intended to support a low-cost improvement to existing technology.

[1] For details, see the description of the IBL Project presented on http://www.ltg.ed.ac.uk/dsea/.

2 Where are backchannels thought to occur?

There are two literatures we can draw on to inspire our model: linguistic theory that predicts where backchannels will occur because of the purpose they serve, and past corpus-based attempts to model backchannel locations.

Theoretically, backchannel continuers will be most interpretable by the speaker if they occur at or before an utterance reaches a *pragmatic completion* — that is, where a segment is "interpretable as a complete conversational action within its specific context" (Du Bois et al., 1993)(p. 147) — but not too early in the utterance. This is because planning the content of an utterance, formulating it, articulating it, and monitoring the partner's understanding are all parallel processes, with monitoring kicking in when planning ends (Levelt, 1998).

Classically, pragmatic completions yield *transition relevance places*, or TRPs for short, where the current hearer can take over the main channel of communication by taking a turn (Sacks et al., 1974), for instance, to clear up something that he does not understand. If the current hearer chooses to take over, then a "turn exchange" is said to occur. If the current hearer chooses not to take over, instead remaining passive or giving feedback through, e.g., a nod, grimace, or backchannel continuer, then the speaker must decide whether to go back or go on. Of course, it is possible for the hearer first to give feedback and subsequently to decide to take a turn. So we would expect speakers to be able to receive backchannel continuers at TRPs, especially when they do not lead to turn exchange, or before TRPs in, say, the second half of their utterance. In their updating of the classic model, Ford and Thompson (1996)(p. 144) describe "complex transition relevance points (cTRPs)" as confluences where intention, intonation, and grammatical structure are all complete. For them, an utterance is grammatically complete if it "could be interpreted as a complete clause . . . with an overt or directly recoverable predicate".

Since speakers can always add phrases after the predicate, grammatical completion is necessary but not sufficient to make a cTRP. Thus linguistic theory suggests that knowing where to find TRPs will help one know where to place backchannel continuers, and that pragmatics, grammar and intonation are all useful cues.

In addition to this theorizing, there have been a number of previous corpus-based studies that have attempted to describe or model the location of backchannel continuers, TRPs, and turn exchanges, by reference to the preceding context. These have tended to concentrate on easy-to-measure phenomena that clearly relate to grammatical and intonational completion: part-of-speech n-grams, pitch, and F0 contour in the immediately preceding context, and pauses at the location itself.

Denny (1985) was concerned with describing the preceding context of only those turn exchanges at which there were pauses of over 65ms, and particularly those at which backchannel continuers occurred. In her description, she considered pitch rise and fall, speaker and auditor gaze, gesture, "filled pauses" such as *mm-hmm*, and grammatical completion.

Koiso et al. (1998), working in a Japanese replication of the same corpus on which our results are based, used all pauses over 100ms as an operational definition of when turn exchange is possible — that is, of TRPs — and considered predictors of whether or not turn exchange occurred at a TRP, and, when it did not, whether or not the hearer produced a backchannel continuer.[2] They used as predictors the immediately preceding part-of-speech plus prosodic features: duration of the final phoneme, F0 contour, peak F0, energy pattern, and peak energy. They found that the best single predictor of either phenomena was the preceding part-of-speech tag, but that combining the prosodic features gave better results, or, preferably, augmenting the part-of-speech tag with the combined prosody features. Turn exchange was indicated by interjections, sentence-final particles, and imperative and conclusive verb forms, together with a rise or fall in intonation. Hearer use of a backchannel continuer was indicated by conjunctive and case/adverbial particles and adverbial verb forms, coupled with the F0 contours flat-fall and rise-fall.

Ward & Tsukahara (2000) modeled the location of backchannel continuers in Japanese and English coversation simply by inserting them wherever the other speaker produced a region of low pitch lasting 110ms. This model is motivated by the observation that such regions often accompany grammatical completion. Their model achieved 18%

[2] The identification of long pauses with TRPs, although understandable in the context of informing work on spoken dialogue systems, is somewhat at odds with previous thinking about turn-taking. Although turn-taking behaviour is culturally dependent , human dialogue is generally considered remarkable for how little silence there can be between turns. A previous study of Map Task data (Bull and Aylett, 1998), bears up Sacks, Schegloff and Jefferson's original (1974) observation that turns often latch, with no perceivable silent gap, or that they even overlap.

accuracy for English and 34% for Japanese.[3]

Although none of these studies is performing exactly the same task as we are, they jointly suggest a range of features that could be included in our model. For example, F0 contour would clearly be useful in predicting backchannel location. However, the challenge of extracting appropriate prosodic features from a pitch tracker lay outside the scope of the research effort reported here. Moreover, the multimodal features considered by Denny seemed too far from the current state-of-the-art in speech recognition systems to be of immediate practical interest. Therefore, for this work, we restrict ourselves to pause duration and part-of-speech tag sequences as inputs.

3 Corpus Analysis

For our modelling, we use the HCRC Map Task Corpus (Anderson et al., 1991),[4] a set of 128 task-oriented dialogues between human speakers of Scottish English, lasting six minutes on average. In half of the conversations the participants could see each others' faces; in the other half, this was prevented by a screen. We ignore this distinction, combining data from the two conditions. Although participants must cooperate to complete the task, their roles are somewhat unbalanced, with one participant, the "instruction Giver", dominating their planning. For this reason, all of our analysis considers where the "instruction Follower" produces backchannel continuers in relation to the instruction Giver's speech.

At the most basic level, a Map Task dialogue represents each participant's behaviour separately as a sequence of time-stamped silences, noises (such as coughing), and speech tokens, to which part-of-speech tagging has been applied. The part-of-speech tag set is based on a version of the Brown Corpus tag set which was modified slightly to better accommodate the corpus ((McKelvie, 2001)). These together allow us to calculate our input features.

We identify Giver TRPs using existing dialogue structure coding. The Map Task Corpus has been segmented by hand into dialogue moves, as described in (Carletta et al., 1997). With the exception of moves in the "acknowledge", "ready", and "align" categories, each move represents one utterance that is either pragmatically complete or, rarely, abandoned. In this system, a ready move is essentially a discourse marker that pre-initiates some larger move, usually an instruction (as in **OK**, *go to the left of the swamp...*), and an align move is usually added to the end of a move to elicit explicit feedback from the partner (as in, *Go to the left of the swamp,* **OK**?). We treat move boundaries as TRPs in our processing, ignoring the two exceptions above which consist predominantly of one-word moves. Failure to remove them affects only our baseline model.

Acknowledgement	Frequency	% of Total
right	1226	29%
okay	587	14%
mm-hmm	462	11%
uh	332	8%
right okay	267	6%
yeah	155	4%
oh right	42	<1%
mm	39	<1%
oh	29	<1%
okay right	19	<1%
aye	17	<1%

Table 1: Frequency of Acknowledgements

The acknowledge move was used to locate backchannel continuers. In this system, all backchannel continuers are acknowledge moves, but not all acknowledge moves are backchannel continuers; following Clark and Schaefer (1991), they include somewhat more substantive ways of moving the conversation forward, such as paraphrasing the speaker's utterance repeating part or all of it verbatim, or accepting its contents. To identify the instruction Follower's backchannel continuers, we filtered the list of their acknowledge moves by removing any that contained content words or words that generally convey acceptance such as *alright*. Table 1 gives the most frequent forms of backchannel continuers resulting from this process, which differ somewhat from Jurafsky et al.'s (1998) analysis of American speech.

4 Description of Models

4.1 Baseline Model

For our baseline model, we planned to insert a backchannel continuer after every n words, for some plausible value of n. This seemed to be the simplest choice in its own right. However, the choice can also be justified as follows. We expect backchannel continuers to be placed at or before intonational phrase boundaries, since these are a primary indicator for TRPs. Spotting these boundaries requires a pitch tracker, but in at least one corpus of spoken English, they are known to occur every five to fifteen syllables (Knowles et al., 1996). We decided to approximate syllables by words. Thus, from each of our Follower backchannels,

[3] Their paper does not specify how these figures are to be interpreted in terms of precision and recall.

[4] The transcriptions and coding for the Map Task Corpus are available from `http://www.hcrc.ed.ac.uk/dialogue/maptask.html`.

we can measure the number of words back to the last Follower backchannel continuer, or Giver TRP, as determined by move boundaries. Figure 1 shows the resulting frequency distribution for the number of Giver words between Follower backchannel continuers.

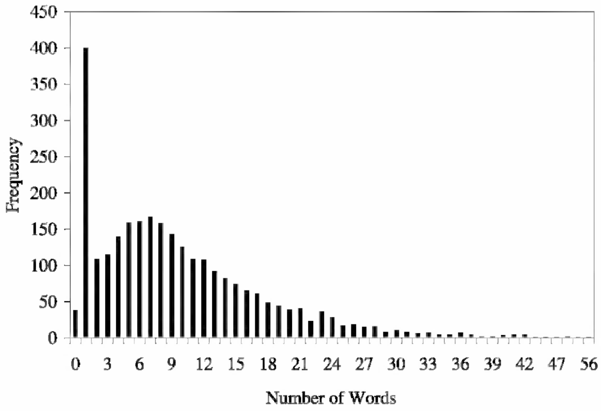


Figure 1: The Number of Giver Words between a Move Boundary and a Backchannel Continuer

In addition to the inclusion of the "ready" and "align" categories (discussed in Section 3), the trigram <s> <aff> <bc> accounts for a continuer occuring after one word. The part-of-speech tag <aff> (affirmative) refers to interjections such as *right*, *okay*, *mm-hmm*, *uh-huh*, *yes* and *no*. Affirmative acknowledgements produced in these circumstances are intended to convey that the Follower has understood the preceding command and is now ready to move on to the next task.

Several models were built that inserted a continuer after n words. The value of n was determined by the frequency of continuers occurring in the data. The variable n increased by one iteration for each model and ranged from four to ten inclusively. The Precision, Recall and F-measure values were found for each model and can be seen in Figure 2. This graph shows all three evaluation metrics for each of the seven models. The smaller the value of n, the more frequently the continuers are inserted. In the model where n equals four, there are 7,147 continuers inserted, but only 3,300 where n equals ten. This is reflected in the recall curve.

The highest F-measure score was produced by predicting a continuer at the mode frequency of every seven words. The score is only 6%.

4.2 Pause Duration Model

Our next model is based simply on pause duration, working from the premise that backchannel continuers often occur at TRPs, and that TRPs often contain

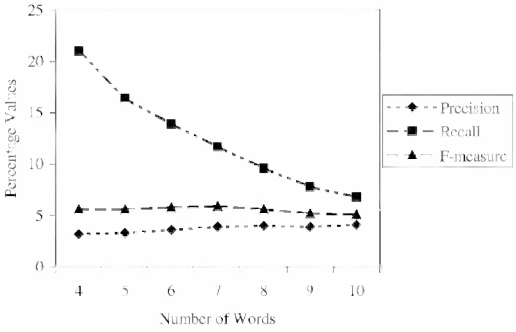


Figure 2: Values for Number of Words

Threshold	Prec.	Recall	F-meas.
0.9	22	59	32
1.0	22	55	31
1.1	22	51	31

Table 2: Highest Performing Pause Duration Models

pauses. As we explained in our discussion of (Koiso et al., 1998), this premise is common, but controversial. Figure 3 compares the durations of the 12% of instruction Giver pauses that overlap with Follower backchannel contributes with the durations of the majority that do not.[5]

Of course, a real-time system cannot wait to see exactly how a long a pause turns out to have been before deciding whether or not to produce a backchannel continuer. In our data, 50% of the pauses lacking backchannel continuers are less than 500ms; moreover, only 11% of pauses this short do attract continuers. For this reason, we apply a threshold; the model works by producing a backchannel for all pauses once they reach a certain length. Eleven models were run, starting with a threshold of below 400ms, and increasing the threshold value in increments of 100ms.

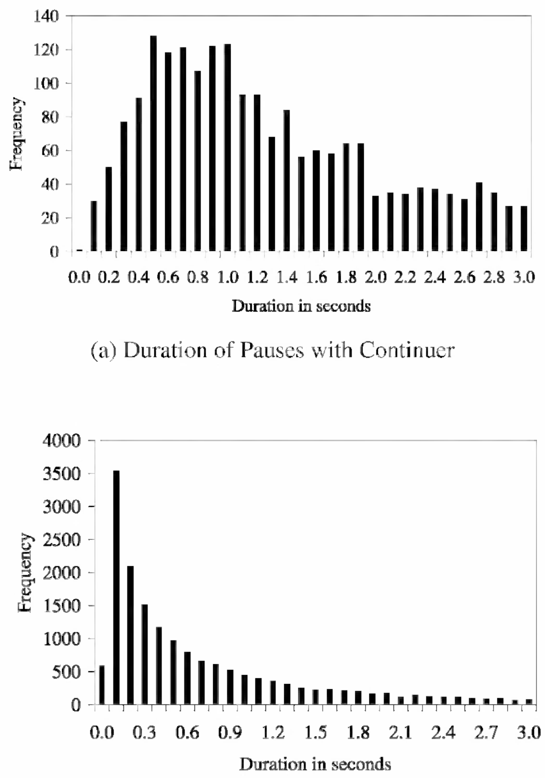
Table 2 shows the values for the highest performing models. The model that only inserts continuers in pauses over 900 milliseconds has the highest overall score. This model was applied to the test set.

4.3 n-gram Part-of-Speech Model

Separating the data into training, validation and test sets was carried out by generating a random dialogue ID. The IDs are in the format q[1-8][en]c[1-8]. A random number was produced for each variable and the files were moved into the relevant directory. The division was approximately 50% training, 30% vali-

[5] For technical reasons to do with the corpus markup, we counted noises that occurred between instruction Giver moves as pauses, but not noises that occurred within moves.

Figure 3: Comparison of Pause Duration



(a) Duration of Pauses with Continuer

(b) Duration of Pauses without Continuers

dation and 20% test data. The validation data was necessary for building the CMU-Cambridge language model, but was concatenated with the training set for the other models.

The model was forced to back-off to a unigram after seeing the continuer tag <bc>, since we did not want this tag to be used as a predictor for any other n-grams. Each move was considered a sentence and given a context tag of <s> and </s> for the start and end of a move respectively, with one move per line. Within the model design the <s> cue automatically causes a forced back-off to a bigram so that the information before the beginning of a sentence is disregarded. This ensured that each sentence was kept as a separate entity; since Follower moves other than acknowledge moves were not represented, sentences were not necessarily in consecutive order.

There are seven occurrences of P(<bc> |) with a back-off value of one. This shows the result of the forced back-off after a continuer tag, and is applied to instances where two continuers appear consecutively. Twenty-one continuers were predicted by the trigram <s> <aff> <bc>. This trigram reflects the manoeuvre "Follower query + Giver affirmative + Follower continuer", discussed in Section 4.1, and accounts for some examples of a continuer occurring after only one word.

Trigram		Probability	Freq./million
P(<bc>	NNS <pau>)	0.263	134.42
P(<bc>	PPO <pau>)	0.220	82.83
P(<bc>	NN <pau>)	0.185	627.64
P(<bc>	PD <pau>)	0.170	33.34
P(<bc>	AP <pau>)	0.150	3.95
P(<bc>	PN <pau>)	0.115	14.66
P(<bc>	RP <pau>)	0.010	113.10
P(<bc>	JJ <pau>)	0.098	25.44
P(<bc>	CC PPG)	0.091	0.74
P(<bc>	DO <pau>)	0.091	4.61

Table 3: Discounted Trigram Frequencies in the CMU-Cambridge Language Model

The ten highest trigram probability counts (using Witten Bell discounting) can be seen in Table 3. The sequence most likely to predict a continuer is a plural noun (NNS) followed by a pause, while sequences consisting of singular noun (NN) plus pause come third. Together, this shows that nouns (either singular or plural) before a pause are good indicators of a backchannel continuer. The tags PPO, PD and PN all represent pronouns and before a pause they make up the second most probable group for predicting a continuer.

A model was built using the three most frequent trigrams as predictors. A second model was constructed using all of the ten most frequent trigrams in Table 3. The aim of this model was to see if increasing the number of factors used in prediction would significantly improve the coverage whilst also maintaining a high accuracy. A continuer was inserted after the occurrence of any of these trigrams in the data.

4.4 Combined Model

The pause duration model was designed to differentiate between pauses that contained continuers and pauses that didn't. Combining the models could be used to filter out the instances where the combination of tags would be more likely to predict an end of move boundary. More precisely a combination of the two models would use the language model to predict the syntactic sequences most likely to determine continuer insertion, and within these, use the pause duration threshold to filter out pauses that are more indicative of an end-of-move boundary.

It is evident from the language model that pause plays an important role in the prediction of continuers. A quarter of all relevant trigrams consist of a part-of-speech tag followed by a pause. This figure includes the most frequent trigrams and those with the highest

	$> 0.6s$		$> 0.9s$	
	three	ten	three	ten
precision	27%	20%	29%	23%
recall	38%	60%	33%	51%
F-measure	32%	30%	31%	32%

Table 4: Comparison of Combined Models

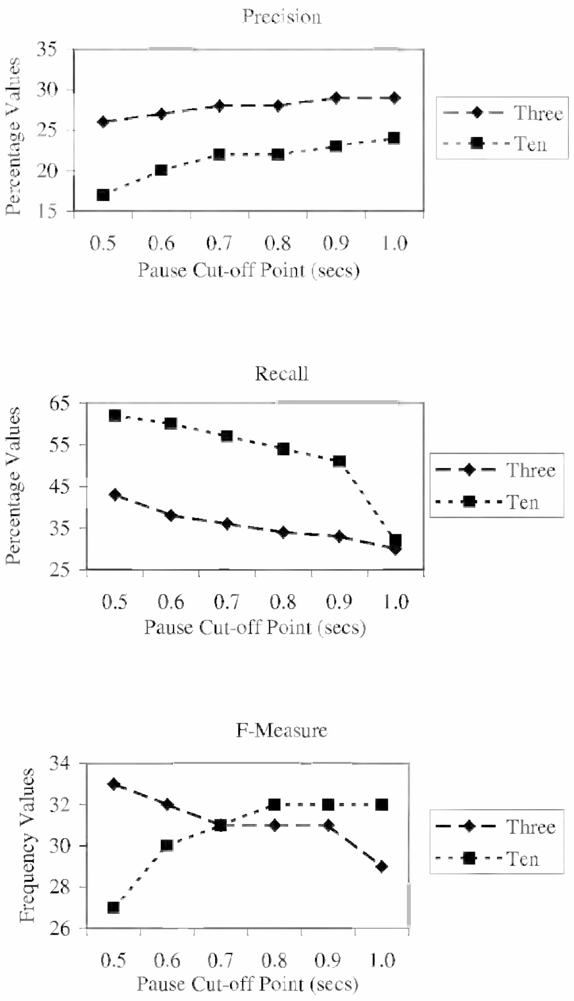
probabilities. Moreover the trigrams that predict continuers are also good predictors of end of move. Using a specified threshold the pause duration model filters out the pauses that are most likely to occur before the end of a move.It could therefore be supposed that combining both the trigram model and the pause duration model should improve the precision and recall. Since this would effectively be cutting out a number of the pauses, a smaller pause duration might be preferable as the higher coverage would compensate for the more concentrated search area. Another way of counterbalancing this effect could be to use the Ten Trigram model, which would increase the number of pauses to which the threshold rule could be applied. A number of combination models were built using both the Top Three Trigram and the Top Ten Trigram models and a pause threshold duration of 500–100ms inclusively. The graphs in Figure 4 show the precision, recall and F-measure results for all the boost models. Graphs A and B demonstrate that the Three Trigram model had consistently higher precision and lower recall scores. Graph C shows that the F-measure values for the Three Trigram models are higher than the Ten Trigram models for the lower threshold values. The values cross at a threshold of 0.7 seconds, after which the Ten Trigram model has the highest F-measure. Finding the ideal compromise between the parameters is difficult to achieve automatically. The F-measure for the Three Trigram model at a threshold of 600 milliseconds is identical to that of the Ten Trigram model at thresholds of 800, 900 and 1000 milliseconds. Using the Ten Trigram model provides the best precision, but the Three Trigram model has a higher recall. For both models the 600 ms threshold has the highest recall, and 900ms the highest precision.

A comparison of these two thresholds can be seen in Table 4. Without carrying out a human evaluation of these models it would be hard to decide between a Three Trigram model with a pause threshold of 600ms and a Ten Trigram model with a threshold of 900ms.

5 Evaluation

The best possible evaluation method, given our aim of low-cost technological improvement, would be to test the acceptability of a dialogue system before and after our models have been incorporated. A potential second best option, having humans judge the naturalness of the models' results independent from a dialogue system, is problematic. Conversational naturalness must be judged in a reasonable amount of left and right-hand context. We could doctor a conversation by excising the real follower's backchannel continuers and re-inserting randomly selected ones where each model predicts, but the results would be judged unnatural because of the knock-on effects on subsequent utterances. A speaker's timings differ depending on whether or not his partner produces a backchannel, and it is difficult to test system insertion of a backchannel where the follower actually produces a more substantive utterance. Thus we have chosen the less explanatory but time-honoured evaluation method of comparing the be-

Figure 4: Comparison of Parameters for the Combined Method



Precision	39%
Recall	36%
F-measure	37%

Table 6: Results of the best model on high backchannel rate data

haviour of our models to what the humans in the corpus do.

One difficulty with evaluating a model such as ours is that human speakers differ markedly in their own backchanneling behaviour. As Ward and Tsukahara (2000) remark, "a rule can predict opportunities, but respondents do not choose to produce back-channel feedback at every opportunity". Because we cannot identify the opportunities that humans pass up, we do the second best thing: cite results both in general and for a relatively high level of backchannel in the corpus. Our reasoning here is that the more backchannels an individual produces, the fewer opportunities they are likely to have passed up.

The models were run on previously unseen test data, the results of which can be seen in Table 5. All models improved on the training models. The baseline model was the worst performer with an F-measure of only 7%. The trigram part-of-speech model and the pause duration models had very similar results, with the pause duration model proving to be a slightly better predictor. The combined model improved the F-measure and importantly the precision. The best model was a five-fold improvement over the baseline.

If we now modify our test set so that it reproduces the behaviour of a speaker with a higher rate of backchannel, we see signficantly improved results. Thus, running the model on the dialogue containing eighty backchannel continuers gives a much higher precision rate, improving upon the best model by 10% as can be seen in Table 6.

5.1 Error Analysis

A number of cases turn up as errors in this evaluation which would not affect the performance of a dialogue system using the model to produce backchannel continuers.

First, the model sometimes posits a backchannel continuer when the route follower actually produces something that has the same effect, but is more substantive (such as a repetition of some of the giver's content). Although the follower's actual utterance provides better evidence of grounding than the system's simple one, modelling the choice of which type of grounding response to produce would be rather tricky for what is likely to be little performance gain.

Second, the model sometimes posits a backchannel continuer when the route follower produces a more substantive, content-ful move. This can be when the follower is not happy for the dialogue to move on, or it can be when the giver has just asked as a question. Of course, a dialogue system using our model would be able to catch these cases because it would know when it wishes to speak, even though by itself, our simple model does not.

Third, a pause was said to contain a backchannel continuer only if the backchannel started or ended within the pause. Instances where the backchannel started slightly *before* the pause would give the trigram *POS* `<bc>` `<pau>`. This particular trigram would not have been found by the language model; after a backchannel backing-off was applied, forcing the language model to count the pause as a unigram. However, after missing this location, the model might well place a backchannel slightly later, during the pause. Changing the location of a backchannel by 500ms does not affect whether or not it was perceived as natural (Ward and Tsukahara, 2000). Thus our evaluation technique overrepresents these misses.

Finally, some of the cases that show up as errors in the evaluation are correct, but the dialogue move coding from which we derived the actual locations of backchannel continuers is not. There is a systematic confusion in our move system between pre-initiating ready moves and acknowledgments (Carletta et al., 1997). These moves share the same realizations, so coders often disagree on which of the two labels to use, especially for the acknowledgments that lack content words and therefore which we counted as backchannel continuers. Even if one accepts the theoretical distinction, a system's behaviour would be perceived as correct if it were to produce something that sounds like a pre-initiator at the same location as a human one, no matter what the system thinks it is.

6 Conclusion

In general there has been very little work carried out on building systems that are capable of placing backchannels. In this paper, we investigated various methods of predicting the placement of backchannel continuers, using only limited processing and information that is readily available to current spoken dialogue systems. Pause duration and a statistical part-of-speech language model were examined. A method combining these two models achieved the best F-measure of 35% and improved on the baseline five-fold. The best previous system (Ward and Tsukahara, 2000) used as its sole predictor regions of low pitch and produced an accuracy of 18% for English.

While our results may not be comparable to other work carried out in the field of natural language pro-

	Baseline	Trigram	Pause	Combined	
				10 Tri + $> .9s$	3 Tri + $> .6s$
Precision	4%	22%	22%	25%	29%
Recall	13%	50%	58%	51%	43%
F-measure	7%	30%	32%	33%	35%

Table 5: Results of the Models on the Test Data

cessing, where scores of 90% or above are not uncommon for tasks such as part-of-speech tagging and statistical parsing, this can be at least partly explained by the fact that humans vary widely in how many of their opportunities for placing a backchannel continuer they actually realize. Our model could potentially be improved by adding words to parts-of-speech in the language model; Ward and Tsukahara (2000) suggest that about half the occurrences of backchannel are elicited by speaker-produced cues. Beyond this, improvements may well require changes to the history that a dialogue system keeps, together with the addition of prosodic information.

References

A. H. Anderson, M. Bader, Bard E. G., E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.

M. C. Bull and M. P. Aylett. 1998. An analysis of the timing of turn-taking in a corpus of goal-orientated dialogue. In R. H. Mannell and J. Robert-Ribes, editors, *Proceedings of ICSLP-98*, volume 4, pages 1175–1178, Sydney, Australia. Australian Speech Science and Technology Association (ASSTA).

J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23:13–31.

H. Clark and E. Schaefer. 1991. Contributing to discourse. *Cognitive Science*, 13:259–294.

R. Denny. 1985. Pragmatically marked and unmarked forms of speaking-turn exchange. In S. Duncan and D. Fiske, editors, *Interaction Structure and Strategy*, pages 135–172. Cambridge University Press.

J. Du Bois, S. Schuetze-Coburn, D. Paolino, and S. Cumming. 1993. Outline of discourse transcription. In J. Edwards and M. Lampert, editors, *Talking Data: Transcription and Coding Methods for Language Research*. Hillsdale.

C. Ford and S. Thompson. 1996. Interactional units in conversation: syntactic, intonational and pragmatic resources for the management of turns. In E. Ochs, E. A. Schegloff, and S. A. Thompson, editors, *Interaction and Grammar*, chapter 3. CUP, Cambridge.

J. Hirasawa, M. Nakano, T. Kawabata, and K. Aikawa. 1999. Effects of system barge-in responses on user impressions. In *Sixth European Conference on Speech Communication and Technology*, volume 3, pages 1391–1394.

D. Jurafsky, E. Shriberg, B. Fox, and T. Curl. 1998. Lexical, prosodic, and syntactic cues for dialog acts. In *ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*.

G. Knowles, A. Wichmann, and P. Alderson, editors. 1996. *Working with Speech: Perspectives on Research into the Lancaster/IBM Spoken English Corpus*. Longman.

H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den. 1998. An analysis of turn-taking and backchannels. *language and Speech*, 23:296–321.

S. Lauria, G. Bugmann, T. Kyriacou, J. Bos, and E. Klein. 2001. Training personal robots using natural language instruction. *IEEE Intelligent Systems*, 16(3):38–45.

L. Lemon, A. Gruenstein, and S. Peters. 2002. Collaborative activities and multi-tasking in dialogue systems. *Traitement Automatique des Langues*, 43(2):131–154.

W. J. M. Levelt. 1998. *Speaking: From Intention to Articulation*. MIT Press, Boston, MA.

D McKelvie. 2001. Part of speech tag set used for MT corpus. Technical report, HCRC. Available from `www.ltg.ed.ac.uk/~amyi/maptask/mt-tag-set.ps`.

H. Sacks, E.A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn taking for conversation. *Language, 50(4)*, pages 696–735.

C. Theobalt, J. Bos, T. Chapman, A. Espinosa-Romero, M. Fraser, G. Hayes, E. Klein, T. Oka, and R. Reeve. 2002. Talking to Godot: Dialogue with a mobile robot. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2002)*, pages 1338–1343.

N. Ward and W. Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 23:1177–1207.

N. Yankelovich, G-A. Levow, and M. Marx. 1995. Designing SpeechActs: Issues in speech user interfaces. In *CHI Conference on Human Factors in Computing Systems*.

V.H. Yngve. 1970. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting, Chicago Linguistic Society*, pages 567–577.