AN ANALYSIS OF THE TIMING OF TURN-TAKING IN A CORPUS OF GOAL-ORIENTED DIALOGUE

Matthew Bull and Matthew Aylett

Human Communication Research Centre, University of Edinburgh

Email: matthew@cogsci.ed.ac.uk

ABSTRACT

This paper presents a context-based analysis of the intervals between different speakers' utterances in a corpus of task-oriented dialogue (the Human Communication Research Centre's *Map Task Corpus*. See Anderson et al. 1991). In the analysis, we assessed the relationship between inter-speaker intervals and various contextual factors, such as the effects of eye contact, the presence of conversational game boundaries, the category of move in an utterance, and the degree of experience with the task in hand.

The results of the analysis indicated that the main factors which gave rise to significant differences in inter-speaker intervals were those which related to decision-making and planning - the greater the amount of planning, the greater the inter-speaker interval. Differences between speakers were also found to be significant, although this effect did not necessarily interact with all other effects. These results provide unique and useful data for the improved effectiveness of dialogue systems.

1. INTRODUCTION

Coversation is one means by which people can coordinate the exchange of information. While the information may be in the form of ideas or facts, the particular characteristic of conversation is that it conveys social information. There are therefore two central considerations: a) how interlocutors convey social signals; b) how interlocutors coordinate these signals. The latter point is of concern here.

The coordination of conversation must deal with the emphasis that interlocutors place on timeliness. A next-speaker (N) may make a contribution as soon as he or she wishes to or is able to, bearing in mind the social signals that a premature contribution will send. If N is too slow in making a contribution when it becomes apparent that there is an opportunity to, the current-speaker (C) may assume that N does not intend to make a contribution (see Sacks, Schegloff, and Jefferson (1974) for their influential model of turn-allocation rules).

The demands of timeliness on N may be great enough that he or she will estimate if an opportunity to make a contribution is imminent. An early account of how this estimate might be achieved (Duncan, 1972) proposes a system of six cues (including prosodic and syntactic elements). More recent accounts (e.g. Ford and Thompson, 1995; Traum and Heeman, 1997) focus on the importance of intonation as a signal of a possible entry point, and posit intonation units as basic units of conversation.

We see therefore that accounts of coordination have tended to focus either on the structure of how interlocutors coordinate their contributions (e.g. Sacks et al., 1974; Clark, 1996), or on the sorts of cues which signal closure of that contribution (Duncan, 1972). Given this, one would expect a detailed and systematic account of the intervals between contributions. Indeed, as indicated above, research into the coordination of conversation relies on the assumption that precision timing is an imperative - that even small differences in timing will either convey some social signal, or affect the success of the coordination process. However, we are aware of only one systematic body of research (e.g. Couper-Kuhlen, 1993). This hypothesises that utterances are coordinated according to rhythmic principles, such that the inter-stress interval at a speaker switch would be equal to (or some integral multiple of) the mean inter-stress interval in the utterance(s) preceding the switch. While this is an appealing theory, as it stands it suffers from being too strong (allowance is made for considerable variability in the ratio of pre-speaker switch to trans-speaker switch inter-stress interval), and there is little compelling evidence in its favour (Bull, 1995; 1998).

In this paper we present an account of the timing of turn-taking based on the assumption that the interval between different speakers' utterances (the *inter-speaker interval*, or *ISI*) is to a large extent determined by the limitations of reaction time and planning time, but also partly by the need to communicate, and to adhere to, a set of social signals. Speaker variations are also significant. We describe the materials on which the hypothesis was tested, the criteria for eliminating inappropriate items from the corpus, and the outcome of the analysis.

2. METHOD

2.1. Corpus

A detailed description of the HCRC Map Task Corpus can be found in Anderson et al. (1991). Essentially, the corpus consists of 128 short dialogues produced during a route communication task. In each dialogue, the two participants had slightly different maps from one another. Each participant worked with a friend and stranger, and each took part in four dialogues - two as Information Giver using the same map, and two as Information Follower using a different map each time. Half the speakers could see their conversational partners; half could not. Each speaker was recorded on a separate DAT channel.

2.2. Coding

Conversational game and move coding (Kowtko, Isard, and Doherty-Sneddon, 1992) was applied to the entire corpus. A move corresponds roughly to a syntactic clause, although its boundaries are set according to functional rather than structural considerations. According to Carletta et al. (1995) a conversational game is "a sequence of moves starting with an initiation and encompassing all moves up until that initiation's purpose is either fulfilled or abandoned." (p.11).

Utterances were defined in terms of move sequencing across speakers. An utterance by a speaker *B* is a sequence of one or more moves, where the start points of the first move (M_{Bi}) and last move (M_{Bj}) in B's utterance (U_B) lie between the start points of two of speaker A's moves (M_{Ai} and M_{Aj} respectively). Note that this definition allows for temporal overlap: M_{Bi} starts after M_{Ai} begins, but not necessarily after M_{Ai} ends. M_{Bj} may continue while A is uttering U_A. However, ISIs lie between the offset of one speaker's utterance, and the onset of the other's. An *exchange* is a pair of utterances surrounding an ISI.

2.3. Data Reduction

The data reduction process can only be described briefly here, and is covered in full detail in Bull (1998). A concern was whether an utterance by speaker B (U_B) could be counted as a response to an utterance by speaker A (U_A). And if U_B were a response, could it be treated as a response to the end of the utterance, or to some earlier part of it? The definition of a response to an utterance is problematic. Generally, we can say that: an utterance U_B is a response to an utterance U_A when U_B is in some way elicited by the content of U_A.

We classified utterances as responses or non-responses using a series of criteria. Some criteria were applied to the corpus automatically to eliminate invalid cases. For example, utterances following backchannelled utterances were eliminated because as defined here, they have no goal-oriented informational content to respond to. Other criteria required some degree of subjective decision-making, and did not necessarily apply in all situations. We isolated a sample of 441 exchanges from the corpus. The intention was then to determine the proportion of response/non-response cases from these, and for each case note one of three durational features.

Excessive overlap - U_B is less likely to be a response to U_A if it starts well before U_A ends. Subjective and automatic criteria were applied to a sample of 236 exchanges. In all cases where the overlap was greater than 1000ms, U_B was not a response to U_A.

Nearly simultaneous onsets - If U_A and U_B start within only a few hundred milliseconds of one another, the delay may be insufficient for B to have interpreted U_A and to have prepared a response. Subjective and automatic criteria were applied to a sample of 126 exchanges. In all cases where the interval between the start times of the utterances was less than 350ms, U_B was not a response to U_A.

Continuations - If one utterance starts very soon after another utterance by the same speaker, the second may in fact be a continuation of the first. Subjective and automatic criteria were applied to a sample of 79 exchanges. In all cases where the interval between the utterances was less than 300ms, U_B was not a response to U_A.

We then used the temporal information relating to each case to set threshold values, which were applied across the whole corpus.

Validation of Exclusion

We carried out a validation study using four other judges. They were presented with a sample (n = 60) of the cases examined by the original judge, J, and a sample of cases not previously judged by J (n = 60). There was poor agreement between each of the four judges and J (maximum K = 0.13). A further analysis revealed that the percentage of adjudged non-responses remaining after application of the cut-offs was: J: 25.2%; Others: 11.7%.. The percentage of adjudged responses eliminated by the cut-offs was: J: 0%; Others: 26.2%. In other words, cut-offs based on J's judgements rather than the other judges' allowed a greater proportion of non-responses and responses in the final data. Accepting J's judgements therefore produced a more conservative, potentially less-biased, data set.

3. RESULTS

Below we list those variables which we found to be significantly related to ISI.

Task Complexity

In the Map Task Corpus, each of the two participants was given a map with a series of features marked on them. However, these features were not necessarily the same. A map in the corpus was classified as +contrast when there was a contrast in the names of the two main features on that map (e.g. *east* lake, and *west* lake). When the giver and follower's maps have the same contrast value, they are +match. Match is therefore an indirect measure of the complexity of the task, because -match maps are more likely to cause the participants difficulties than +match maps.

The mean ISI of +match exchanges was significantly lower than the mean ISI of -match exchanges (+match: mean = 469ms, n = 5816, s.d. = 718ms; -match: mean = 522ms, n = 5201, s.d. = 819ms; t = 3.66, d.f. = 11015, p = 0.0002). We may conclude from this that a mismatch in the main features on a map result in greater mean ISIs because the participants must spend added time between utterances solving their problem.

Task Familiarity

Figure 1 below shows a rough downward trend in the mean ISI from the first dialogue in a quad, to the eighth. This indicates that familiarity with the Map Task may be significantly linked to ISI, since with each successive conversation number each participant would have been more familiar with the general task involved. Without such strategies, it is likely that more planning would be required and mean ISIs would be greater.

An analysis indicated that the ordering of dialogues played a

significant role in the determination of ISI duration. The mean ISIs for dialogues 1 and 2 were longer than for dialogues 3-8 (1-2: mean=561ms, s.d. = 871ms, n = 3140; 3-8: mean = 467ms, s.d. = 721, n = 7877; t = -5.78, d.f. = 11015, p < 0.0001). There was therefore good evidence that lack of familiarity in a task is reflected in increased ISI duration, possibly caused by the need for greater planning and decision time by both participants.

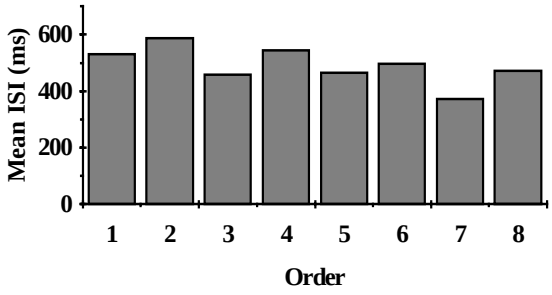

Figure 1: Mean ISIs for each conversation number in order.

Figure 1 also shows a 'stepping' pattern between odd and even dialogue numbers. The odd-numbered dialogues were all +match, and the even-numbered dialogues were all -match. Each of the odd-numbered dialogues were found to have a lower mean ISI than its following even-numbered dialogues had, supporting the match effect mentioned above.

Conversational Game Boundary

If the introduction of a new game depends on some extra decision time or planning time (for example see Clark, 1996) then ISIs should be longer between utterances separated by a game boundary than between utterances within a game. The mean ISI at game-boundary exchanges was significantly greater than the mean for exchanges elsewhere (game boundary: mean = 621ms, s.d. = 775ms, n = 972; elsewhere: mean = 431ms, s.d. = 726ms, n = 7927; t = -13.92, d.f. = 11015, p < 0.0001).

Intervals at game boundaries might be longer than elsewhere because their beginnings are characterised by certain move categories which are not found elsewhere. We therefore carried out a repeated measures ANOVA, which used equal group sizes for each of the move categories at both game boundary and elsewhere locations (in fact five of the twelve move categories were omitted because of a small sample size). This crossed game boundary ISIs and elsewhere ISIs with move category, and found that there was a significant game boundary effect (F (1, 1180) = 24.65, p < 0.0001), and move category effect (F (5, 1180) = 3.80, p = 0.002). Importantly, the interaction between the two was significant (F = 3.3 (5, 1180), p = 0.0057), indicating that ISIs at game boundaries are to some extent dependent on move categories. This supports the notion that game boundary effects are partly the result of the patterns of move category found only at game boundaries. But the results show also that there is a significant separate game boundary effect.

Role

Each participant was assigned the role of either Instruction Giver or Instruction Follower. We supposed that there would be a difference in ISI duration according to whether there was a switch from giver to follower or vice versa, because of the different planning and decision requirements of each speaker. An instruction giver may be required to plan ahead more than an instruction follower, and to develop strategies for conveying information as efficiently as possible. Consequently, ISIs preceding a giver's utterance may be longer than those preceding a follower's utterance.

However, mean ISI is significantly greater when it falls across a giver-follower speaker switch than when it falls across a follower-giver switch (giver-follower: mean = 541.4 ms, s.d. = 824.4 ms, n = 6798; follower-giver: mean = 417.6 ms, s.d. = 658.8 ms, n = 4219; t = -8.26, d.f. = 11015, p < 0.0001).

Further analyses revealed that the importance of the role variable does not lie in factors such as game boundary. Mean intervals are consistently greater for giver-follower exchanges than for follower-giver exchanges, irrespective of whether the exchange occurs across a game boundary, or within a game, as shown in Figure 2 below. The general conclusion was that instruction followers are more likely to leave a long ISI before speaking than instruction givers, possibly as the result of silences taking place when tasks are actually being carried out (e.g. drawing a line on the map).

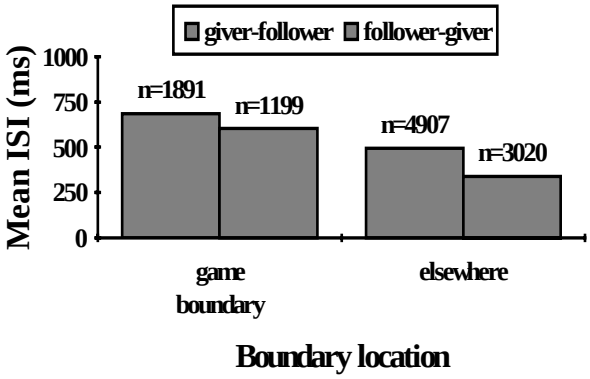

Figure 2: Mean ISIs according to location and role of speaker. F (1, 11013) = 44.65, p < 0.0001

Eye contact

This variable is a measure of whether the two participants were able to see each other, and not whether they were actually looking at each other. The corpus was split into eye contact dialogues and non eye contact dialogues. The mean ISI for eye contact dialogues was significantly higher than for non eye contact dialogues (eye: mean = 579.7ms, s.d. = 669.5ms, n = 5995; non eye: mean = 422.2ms, s.d. = 862.6ms, n = 5022; t = -10.78, d.f. = 11015, p < 0.0001).

This shows that the potential to see the other participant does make a difference to mean ISI. This may be because being able

to see the other participant permits a greater tolerance of longer ISIs than when there is no possibility of seeing the other speaker. Temporal coordination need not be as tight when participants are able to see each other. When it is not possible to see a partner in a conversation, mean ISI may be lower because interlocutors generally over-compensate, and become less tolerant of longer ISIs.

Speaker differences

We maintained that the personal characteristics of the participants in the Map Task study could affect distributions of ISIs. We tested specifically for a relationship between mean ISI and the identity of the speaker making the second utterance in an exchange pair (the speaker having a direct influence on the ISI). We found a significant speaker effect (F (63, 10953) = 10.96, $p < 0.0001$). We carried out a series of ANOVAs to test for interactions between this variable and others (although because of the design of the experiment this was limited by impossible interactions and small cell sizes in some interactions.

Speaker identity was found to be significantly related order effects (speaker - F (59, 10311) = 9.23, p < 0.0001; order - F (3, 10311) = 11.26, p < 0.0001; interaction - F (177, 10311) = 3.03, p < 0.0001). However, a separate ANOVA crossing speaker identity, game boundary, and role variables (speaker - F (59, 10515) = 8.31, p < 0.0001; game - F (1, 10515) = 144.99, p < 0.0001; role - F (1, 10515) = 38.44, p < 0.0001) found an interaction between speaker identity and role (F (59, 10515) = 2.68, p < 0.0001), but not between speaker identity and game boundary. In other words, we can conclude that game effects probably act largely independently of speaker differences.

4. CONCLUSIONS

Important factors in the determination of ISIs can be grouped broadly into three types: a) speaker differences; b) factors set by the conditions of the task, such as the potential for interlocutors to see each other, their respective assigned role in the task, and the design of the maps; c) factors concerned with dialogue structure, such as the effect of game boundaries and different move categories.

First, we have demonstrated here that dialogue structure is significantly related to ISI duration. In fact, to take these results one step further one could claim that this data should aid considerably models which calculate the likely dialogue structures of conversations. However, it is still not clear to what extent the game boundary effect is the result of processing or planning time, or the result of acting as some form of a 'task finished' signal.

Second, we have demonstrated that factors specific to a task are related to ISI duration - for example task complexity, and learning effects. These factors in turn reflect the limitations of reaction time, processing time, and planning time. Other task-specific variables (eye contact and role) are somewhat more problematic. The eye contact effect is possibly the result of greater overlap in non eye contact situations because important gestural turn-closure signals are missing. The role effect may result from differences in the sorts of task implicitly given to giver and follower in the experiment.

Finally, we have shown that while speaker differences are significant, they do not interact with the significant effects of dialogue structure. This lends further weight to the significance of the relationship between dialogue structure and ISI.

5. REFERENCES

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

2. Bull, M.C. "An appraisal of rhythm as a coordinator of turn-taking", *Proceedings of the XIIIth International Congress of Phonetic Sciences. Vol. 3, 1995,* 480-483.

3. Bull, M.C. *The Timing and Coordination of Turn-taking*. Unpublished doctoral thesis, University of Edinburgh, 1998.

4. Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. and Anderson, A. "The coding of dialogue structure in a corpus". In J. A. Andernach, S. P. van de Burgt and G. F. van der Hoeven, (eds.), *Proceedings of the Ninth Twente Workshop on Language Technology: corpus-based approaches to dialogue modelling*, 1995, 25-34.

5. Clark, H.H. "Using Language", Cambridge University Press, Cambridge, 1996.

6. Couper-Kuhlen, E. "English speech rhythm: form and function in everyday verbal interactions", John Benjamins, Amsterdam, 1993.

7. Duncan, S. "Some signals and rules for taking speaking turns in conversation", *Journal of Personality and Social Psychology, Vol. 23, 1972,* 283-292.

8. Ford, C.E. and Thompson, S.A. "Interactional Units in Conversation: syntactic, intonational, and pragmatic resources for the management of turns". In E. Ochs, E.A. Schegloff and S.A. Thompson (eds.), *Interaction and Grammar*. Cambridge University Press, Cambridge, 1995.

9. Kowtko, J. C., Isard, S. D. and Doherty-Sneddon, G. M. "Conversational Games within Dialogue", *Research Paper HCRC/RP-31*, Human Communication Research Centre, Edinburgh, 1992.

10. Sacks, H., Schegloff, E.A. and Jefferson, G. "A simplest systematics for the organization of turn-taking for conversation", *Language, Vol. 50, 1974,* 696-735.

11. Traum, D., and Heeman, P. "Utterance units in spoken dialogue". To appear in E. Maier, M. Mast, S. LuperFoy (eds.), *Processing in Spoken Language Systems*. Springer Verlag, Heidelberg, 1997.