Let's You Do That: Enquiries into the Cognitive Burdens of Dialogue

E. G. Bard,
TAAL, HCRC
U. of Edinburgh
Edinburgh EH8 9LL
ellen@ling.ed.ac.uk

A. H. Anderson
Dept of Psychology, HCRC
U. of Glasgow
Glasgow G12 8QB
anne@psy.gla.ac.uk

Y. Chen
TAAL, HCRC
U. of Edinburgh
Edinburgh EH8 9LL
yiya.Chen@let.ru.nl

H. Nicholson
TAAL, HCRC
U. of Edinburgh
Edinburgh EH8 9LL
hannele@ling.ed.ac.uk

C. Havard
Dept of Psychology, HCRC
U. of Glasgow
Glasgow G12 8QB
c.havard@psy.gla.ac.uk

Abstract

Most discussions of audience design assume that it rests on speakers' uptake of information about listeners' knowledge. The cognitive difficulty hypothesis (Horton and Gerrig, 2004 in press a) proposes that speakers provide less tailored design when the cognitive cost of uptake or recall increases. Yet the principle of mutual responsibility implies that cognitive load should to be shared efficiently: listeners should provide information which would be difficult for speakers to discover themselves. Two map task experiments examine speakers' uptake of information about listeners' knowledge and their responses to listeners' difficulties. Both experiments show that uptake is poor where it would be most useful: speakers attend very little to feedback in the form of simulated listener eye-tracks which directly indicate discrepancies between participants' knowledge. The second experiment shows that verbal feedback, though harder to interpret than gaze, generates more helpful responses in the form of Dialogue Transactions which correct listener errors and in the form of Game Moves which focus on listener knowledge. We propose that the instructor's priority is relating her own knowledge and that she will be deflected only when overtly called on to acknowledge a discrepancy between her own knowledge and the listener's.

1 Introduction

Recently experimental psycholinguists have given a great deal of attention to dialogue, with particular emphasis on the extent to which speakers design utterances for the benefit of their interlocutors. Audience design of this kind is taken to validate the notion of common ground in a psychological model of the process of conducting dialogues: if speakers maintain a model of their interlocutor's knowledge as well as their own, the intersection, the knowledge held in common, can be estimated[1]. A parallel line of research addresses common

[1] Strictly, common ground is only that shared knowledge which is mutually acknowledged as shared (Clark & Marshall, 1981; Barr & Keysar, 2004). We deal here with shared knowledge, both because it is usually what is at stake in the experimental literature and because it appears to be central to the view that we develop.

Proceedings of the 9th Workshop on the Semantics and Pragmatics of Dialogue, June 9-11, 2005, Nancy, France.

ground from the listeners' perspective, examining how a listener's knowledge about the what the speaker knows can affect that listener's interpretation of the speaker's referring expressions.

These experiments are based on several predictions involving the notion of common ground. The first gives every speaker responsibility for discovering what information is in common ground. To do this, it is predicted, each must at least attend to clues to the other's knowledge (Clark & Carlson, 1982, Clark & Krych, 2004). The second requires each speaker to exploit these cues when framing her own utterances. The third invokes the theory of mind in interpretation: it predicts that, as a listener, any interlocutor will consider only those candidate referents which he knows to be in or derivable from knowledge held in common.

Underlying this research is the assumption that common ground, the knowledge held mutually, will be established when each interlocutor performs two tasks: modeling the other's knowledge and maintaining her own. Clearly, one of these record keeping tasks is easier than the other: a participant's own experience can be recorded in episodic memory and can function via computationally inexpensive associative processes like priming (Pickering & Garrod, 2004) or resonance (Horton & Gerrig, in press b). The upkeep for a model of the interlocutor's knowledge can be much more costly (Bard & Aylett, 2004; Carletta & Mellish, 1996; Pickering & Garrod, 2004) and may involve chains of inferences about the interlocutor's actions, intentions or conceptions (Clark and Marshall, 1981). For this reason, dialogue is a joint project, a game for two players which can best be played if each player makes the contributions that keep the other player's task feasible. Though the principle of least collaborative effort (Clark & Wilkes-Gibbs, 1986) allows both players to make gradual contributions to the establishment of common ground, it is possible to go a step further.

Studies of audience design take the notion of joint responsibility for creating common ground to mean that each participant has full responsibility for maintaining and embellishing the models of both speakers. In many ways, this amounts to cost-duplication. The principle of least collaborative effort means that joint responsibility should be a kind of cost sharing, with players assuming not identical, but complimentary responsibilities (Carletta & Mellish, 1996). Each should attend to his or her own knowledge and present it to the other when necessary. In this view of joint responsibility, no interlocutor need be responsible for information which the other can provide more economically. This latter interpretation of joint responsibility seems to come close to Clark, Schreuder and Buttrick's (1983) definition of optimal design.

Thus, audience design, in the sense of adjusting one's contributions to what the interlocutor knows, is not an absolute requirement; nor is listener modeling principally the responsibility of the speaker. Instead, speakers can design their utterances as suits their current personal knowledge or the currently known common knowledge, without actively seeking additional detail about the listeners. It is the their listeners' responsibility to provide them with indications of their own share of common ground, drawing on cheap and cheerful own-knowledge record keeping. The Monitor and Adjust model of dialogue (Horton & Keysar, 1996), under which speakers monitor both their own output and their interlocutor's feedback, is similar in spirit. It makes slightly stronger assumptions about self-monitoring than this position does, and it follows Clark and Schaefer in concentrating on listeners' rejection particular utterances, rather on their own contribution to common ground.

In summary, then, the theory of dialogue as joint activity makes contradictory predictions. Where joint responsibility is duplicated responsibility, speaker A is responsible for tracking speaker B's knowledge. Where joint responsibility is

shared responsibility, B is responsible for revealing his pertinent knowledge to A. There is evidence for both positions.

On the one hand, Speakers monitor listeners' activity and gestures while speaking (Clark & Krych, 2004). Speakers maintain forms of referring expression with a particular interlocutor (Brennan & Clark, 1996) and are disrupted if that interlocutor chooses a different expression (Metzing & Brennan, 2003). Speakers initially provide more detail in description, particularly atypical detail, for listeners who cannot see the picture described (Lockridge & Brennan, 2002). Speakers incrementally supply descriptive phrases in the order in which they can most conveniently be used by listeners (Haywood, 2004). Listeners will interpret referring expressions as if addressed to them (Hanna, Tanenhaus, & Trueswell, 2003).

On the other hand, speakers may provide egocentric descriptions initially and audience-related descriptions somewhat later (Dell & Brown, 1991); habitually utter syntactically ambiguous structures, where unambiguous paraphrases are available (Ferreira & Dell, 2000); describe objects when under time pressure in ways which are unhelpful to listeners (Horton & Keysar, 1996); perform faster production adjustments egocentrically and slower ones with only modest care for the listener (Bard et al., 2000; Bard & Aylett, 2004); interpret referring expressions as naming objects salient to themselves but patently unknown to the speaker (Keysar, Lin, & Barr, 2003); require experience as an addressee in an object selection task before providing evidence of audience design in their own utterances (Haywood 2004).

To deal with these contradictions, Horton and Gerrig (2004 in press a) have recently proposed a difficulty model of common ground construction, under which listener modelling is subject to effects of the cognitive effort involved. Modelling will be slow or less complete when it is more difficult. Horton and Gerrig show that interlocutors adhere more closely to principals of audience design in a later dialogue when it is simpler to distinguish their co-participants in terms of the task pursued in an earlier dialogue.

The present paper asks whether audience design, cognitive difficulty, or joint responsibility controls behaviour in dialogue. We investigate this question in a route communication task where two variables affect cognitive load. One is the source and specificity of the information about the listener's knowledge state. One source is the direction of the listener's gaze. If A says "Go to the large oak tree" and sees B looking at the bridge instead, little inference is needed to devise a correction (get B from the bridge to the oak tree). The other source is typical verbal feedback. If, when told to go to the large oak, B replies 'Don't follow', A will not know whether B lacks the oak, has two small oaks, or cannot understand the instruction. Even 'Don't have it' could mask a mismatch of map landmarks or a misunderstanding of instructions. A chain of inferences and investigations are required to tailor a solution to the listener's problem. The second measure of cognitive load, time pressure, is used because remarkably egocentric behaviour can occur when speakers are pressed to respond quickly (e.g., Horton & Keysar, 1996), and much better audience design when they respond at leisure.

Listener modeling, the difficulty model, and joint responsibility make different predictions here. An assiduous modeler of common ground will attend to all sources of information about the listener's knowledge: a speaker who says 'Fine!' but is looking in the wrong place needs to be told that he has a problem. One who says 'Can't see it' and is apparently lookiing in the right place needs a different kind of help. This attention should be maintained as long as time pressure permits competent dialogues to be completed. If common ground is cultivated more when the cost of cultivation is less, speakers should attend

to visual feedback at least as assiduously as to verbal replies (Clark & Krych, 2004; Pomplun et al, 1997), should give proportionately more attention to feedback when unhurried than when rushed. Joint responsibility, however, predicts that processing cost, uptake, and audience design are not related. Instead, because listeners' verbal contributions to the construction of common ground are the key to joint action in dialogue, speakers may habitually ignore visual feedback and attend instead to their interlocutors' explicit demands.

2 Experiment 1

Experiment 1 tests the ability of visual feedback alone to supply the role of the listener. The direction of the interlocutor's gaze is an important component of co-presence. The ability to see where the interlocutor is looking greatly enhances the utility of virtual co-presence (Gale & Monk, 2000). Here, visual feedback is instantiated a way that allows us to determine when it is attended to: the simulated eye-track of a distant listener is projected onto the monitor showing the route which the participant describes and the participant's genuine eye-track is examined for time spent looking at the interlocutor's track.

2.1 Method

Materials: Four different maps of fictitious locations each included a route defined by a number of labeled cartoon landmarks. Eight or 9 route-critical landmarks were designated *correct* and 4 non-adjacent items were to be *missed*. Other, irrelevant landmarks assured that the Instruction Giver (hereafter 'IG') always had to distinguish a route-critical landmark from a number of others. To simulate the gaze of an Instruction Follower (hereafter 'IF'), a red square was superimposed on a sequence of landmarks with saccades of random length and direction outward from each fixation target. For *correct* landmarks the fixation target was the route-critical landmark itself. For *missed* landmarks, the fixation target was a *wrong* landmark,elsewhere on the map. Though target sequence was preprogrammed, migration was initiated by the experimenter as soon as the participant named the next route-critical landmark. To create a usesable trial response, the experimenter had to advance the IF feedback square between the IG's instruction to move to the new landmark and any instruction to correct the IF or to move to the following landmark. The feedback square returned to the route after a detour only when the participant gave the appropriate instructions or moved advanced to the next landmark on the route.

Apparatus: Participants viewed maps on a flat screen monitor at a distance of 60 cm. Eye movements were recorded on an SMI remote eye-tracking device placed on a table below the monitor and using Iview version 2 software. Speech was recorded in mono using Asden HS35s headphone /microphone combination headsets. Video signals from the eye tracker and the participant monitor were combined.

Design: All participants served as IG for all 4 maps, with 2 under a 1-minute time limit and 2 without limit. One map in each time pressure condition included visual feedback. The Time Pressure and Feedback combinations were applied to maps by Latin Square.

Procedure: Participants were met with a confederate and asked to take the role of IG while the confederate worked as IF in another room. IGs were asked to describe the route on each map to the IF so that the latter could reproduce it by using her mouse to traverse a similar screen displaying a similar but not identical map. The feedback and timing conditions were explained and announced before each trial. Participants were fully debriefed. None suspected the true nature of the experiment.

Participants were Glasgow University students (aged 17-24), all with normal or corrected to normal vision, all native English speakers, and all paid £5 for partici-

pating. Participants were rejected from the final set if eye-tracking capture fell below 80% of experiment time on any map or if the experimenter missed the critical time-window for moving the IF feedback square for any wrong item on a map or for more than 30% of the correct items. Testing continued until 24 participants passed these criteria and filled a balanced design.

2.2 Results

Interactive behaviour: To discover whether participants engage in something other than monologue with purely visual feedback, we coded their transcribed speech as Transactions and Conversational Game Moves (Carletta et al., 1997). A Transaction is as section of a dialogue which achieves an identifiable subgoal of a non-linguistic task. A new type of Transaction, a Retrieval, was identified, in which IG explicitly directed a lost IF back to the route. ANOVAs were calculated by subjects (F_1) and/or by items (F_2) as appropriate. Absent in the no feedback condition, Retrievals were found in the trials with visual feedback. The usual route-advancing 'Normal' Transactions accordingly fell in frequency between no feedback and visual feedback conditions (Feedback: $F_1(1,23) = 24.68$, $p < .001$). Conversational Game Moves are stages of the linguistic task which manipulate information and common ground. Moves which were specifically interactive in that they would not be expected in monologue (queries, aligns, acknowledges) were significantly more common with visual feedback than without (F_1 $(1,23) = 21.48$, $p < .001$). Time pressure affected only gross length of dialogues and amounts of gaze.

Attention to interlocutor knowledge: As the participants' speech had become more like dialogue in the feedback condition, listener modeling ought to be encouraging good uptake of cues to listener knowledge. Both where the 'IF's' gaze rested on the correct landmark and particularly where it digressed to an off-route landmark, IG should look longer at the targeted landmark than in the control condition, which lacked feedback. IG should look less at the route-critical landmarks which the IF missed, because her attention should be diverted to the landmark where IF appeared to be mistakenly gazing. In fact, a very different pattern emerged from mean total time spent gazing in the region of a landmark. As predicted, IGs looked longer at landmarks attracting correct IF gaze than at the same landmarks in the no-feedback condition, an average increase of 1.4 sec per landmark. Contrary to prediction, IG also looked significantly longer at on-route landmarks which IF missed (610 msec) but not at the distant 'wrong' landmarks under IF's gaze (430msec) (Landmark type: $F_1(2,46) = 4.10$, $p = .023$; Correct v wrong, $p < .05$). Relatively little time, then, was absorbed by attending to discrepancies between IG's and IF's knowledge. Instead, participants gazed at the on-route landmarks, whether IF's attention was directed to them ('correct landmarks') or not ('missed landmarks').

3 Experiment 2

Experiment 2 tests ease of absorbing listener-knowledge by comparing verbal to visual feedback.

3.1 Method

Materials comprised 6 maps, 4 derived from those used in Experiment 1 and 2 created to the same model.

Design: All participants used all 6 maps, 2 maps in each of three feedback conditions: no feedback, single channel (verbal for Group A participants, visual for those in Group B), and dual channel (verbal and visual). One trial in each modality condition was performed within a time limit of 2 minutes, while the other had no time limit. The order of feedback conditions was as just described in each time pressure condition. The order of time pressure conditions and the assignment of maps to condition were counterbalanced over the design.

Feedback was delivered once the subject had introduced each route-critical landmark. Verbal feedback was provided by the confederate according to a script: in negative replies, the confederate claimed not to be able to see the landmark, or follow the instruction, but did not explicitly describe any guess, location, or difficulty. Visual feedback was delivered as in Experiment 1. On each map, 7 to 9 route-critical landmarks received correct visual feedback with (concordant) positive verbal feedback (the IF 'gaze' was on the correct landmark and the IF said that it was); 3 landmarks had correct visual feedback and (discordant) negative verbal feedback, 3 had wrong visual feedback and (discordant) positive verbal feedback, 3 had wrong visual feedback and (concordant) negative verbal feedback.

Procedure and apparatus were as for the Experiment 1, with the addition of the extra conditions. Again, subjects were debriefed and the two participants who were suspicious about the true nature of the experiment were replaced.

Participants were 36 Glasgow University undergraduates, 18 per group, each paid £5. An additional 13 participants had been replaced because one of their 6 trials fell below the eye capture criterion.

3.2 Results

Results were coded as in Experiment 1 with the additions of new conditions.

Attention to interlocutor knowledge: Assiduous listener modeling regardless of difficulty would demand that IG track IF's gaze as well as attending to IF's verbal feedback**.** If difficulty discovering pertinent listener knowledge is critical, IG should track simpler visual information, especially where the IF's and IG's interpretations apparently diverge, and the feedback square moves to the wrong landmark. When visual and verbal feedback disagree (as they do in discordant conditions) visual feedback should take precedence: speakers should look at the wrong landmarks, even if verbal feedback is positive. Timed dialogues should show proportionately less attention to IF-only information. If, however, it is not the speaker's task to track the listener's knowledge, there is no particular attraction in divergent listener gaze.

The effects of both feedback and time pressure showed this last pattern. In total, IGs looked less at landmarks under time pressure (F_1 (1, 34) = 48.08, $p < 0.001$), but the reduction applied to all landmarks *except* the IF-specific wrong landmarks ($F_1(3, 116) = 13.83$, $p < 0.001$; $F_2(5, 126) = 11.773$, $p < .001$) where gaze durations were minimal throughout (< .295sec vs 5 to 6sec for all others). IGs looked longer as each feedback channel was added (F_1 (2, 59) = 329.95, $p < 0.001$), but feedback did not prolong gaze on the wrong landmarks ($F_2(10, 244) = 3.26$, $p < 0.01$). We checked Transactions including feedback for any examples of speaker gaze at listener position, no matter how brief. Table 1 shows that speakers more often than not (59% of trials) failed to look at the Follower feedback square at all when it targeted the wrong landmark, though they more often looked at it (55%) when it targeted the correct landmark which they were in the course of describing ($F_1(1,34) = 7.70$, $p = .009$).

Verbal feedback	Visual Feedback	
	Correct	Wrong
Positive	.51	*.45*
Negative	*.59*	.37

Figure 1. Proportion of feedback episodes attracting speaker gaze to feedback square: Effects of combinations of visual and verbal feedback in dual channel conditions (Italics represent discordant feedback)

Interactive behaviour: If speakers always tailor their output to interlocutors, any feedback should encourage interactive behaviour which solves listener problems. If cognitive difficulty affects audience design, then interactive contributions should be more common when speakers can ac-

cess simply processed visual indications of the listener's knowledge. If verbal feedback is key, as joint responsibility suggests, then it should attract interactive talk more than visual cues do. As Figure 2 below reveals, the third pattern holds.

Retrieval Transactions, which bring errant Followers back from places that can be seen with visual feedback, are far less common with unambiguous visual feedback alone (7% of opportunities) than with only ambiguous verbal feedback or with visual and verbal information that may conflict (27%) ($F_1(1,34) = 90.80$, $p < .001$, cell comparisons at $p < .05$). A similar pattern is found for Interactive Moves (6% v 30%: $F_1(2,68) = 36.53$, $p < .001$).

Dependent variable	Single channel	Feedback channels 0	1	2
Retrieval transactions	Verbal		.27	.27
	Visual		.07	.27
Interactive moves	Verbal	.00	.31	.34
	Visual	.01	.06	.25

Figure 2. Effects on rate of interactive behaviours from feedback channels and modality of single-channel condition

4 Conclusion

Two experiments have shown, first, that visual feedback alone can make speakers' instructions more like a dialogue and, second, that speakers did not pay close attention to direct visual evidence for their listener's problems, however simple it might have been to interpret. In fact, they took up this feedback only when it fell on route-critical landmarks which they were already fixating in order to describe the route. They avoided looking at the spots where their listener's gaze had mistakenly focused. In the second experiment, gaze showed a similar pattern. Moreover, though visual feedback gave clear evidence for the location of the lost IF, it was not sufficient to launch a rescue: both interactive Moves and Retrievals depended on the presence of verbal feedback either alone or in combination with visual.

The results do not sit comfortably with a model which demands continuous uptake of listener information. Nor do they show the responses to time pressure or ambiguity that might support a cognitive load model. Instead the results point to joint responsibility: Verbal feedback seems to be required to draw participants' attention to the problems at hand. Perhaps verbal feedback has this quality because an intentional signal of distress is needed to derail IG's inadequate descriptions. Or perhaps visual feedback is ignored because IG can simply wait for the IF to re-appear without knowing where or how he is lost. Clearly in this paradigm, responsibility for designing adequate instructions was jointly held.

Acknowledgments

This work was supported by EPSRC (UK) grant GR/R59038/01 to E. G. Bard and grant GR/R59021/01 to A. H. Anderson. Dr. Chen is now at Radboud Universiteit, Nijmegen.

References

Bard, E. G., Anderson, A., Sotillo, C., Aylett, M. Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *JML, 42*, 1-22.

Bard, E. G., & Aylett, M. P. (2004). Referential form, word duration, and modeling the listener in spoken dialogue. In J. Trueswell and M. Tanenhaus (eds.), *Approaches to studying world-situated language use*. Cambridge, MA: MIT Press, pp. 173-191

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *JEP: LMC, 11*, 1482-1493.

Carletta, J., & Mellish, C. (1996). Risk-taking and recovery in task-oriented dialogue. *J Pragmatics, 26,* 71–107

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. & Anderson, A. 1997. The reliability of a dialogue structure coding scheme. *Comput Linguist, 23*, 13-32

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149). Washington, DC: APA

Clark, H. H., & Carlson, T. B. (1982). Hearers and speech acts. *Language, 58,* 332-373.

Clark, H. H., & Krych, M.A. (2004). Speaking while monitoring addressees for understanding. *JML, 50,* 62–8

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. Webber, & I. Sag (Eds.), *Elements of discourse understanding* (pp. 111–222). Cambridge: Cambridge University Press.

Clark, H. & Schaefer, E. (1987). Collaborating on contributions to conversations. *Lang Cognitive Proc, 2*, 19-41.

Clark, H. & Schaefer, E. (1989). Contributing to discourse. *Cognitive Sci, 13*, 259-294.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground and the understanding of demonstrative reference. *JVLVB, 22*, 245-258.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22,* 1-39.

Dell, G., & Brown, P. (1991). Mechanisms for listener-adaptation in language production: Limiting the role of the "model of the listener". In D. J. Napoli & J. A. Kegl (Eds.), *Bridges between psychology and linguistics: A Swarthmore Festschrift for Lila Gleitman* (pp. 111–222). Hillsdale, NJ: Erlbaum.

Ferreira, V., & Dell, G. (2000). Effect of ambiguity and lexical availability on syntactic ambiguity on syntactic and lexical production. *Cognitive Psychol, 40,* 296-340.

Gale, C., & Monk, A. F. (2000). Where am I looking? The accuracy of video-mediated gaze awareness. *Percept Psychophys, 62*, 586–595.

Hanna J. E., Tanenhaus, M. K. & Trueswell, J. C. (2003) The Effects of common ground and perspective on domains of referential interpretation. *JML, 49*, 43-61

Haywood, S. (2004). *Optimal design in language production.* Ph.D. Dissertation. University of Edinburgh.

Horton, W.S., & Gerrig, R. J. (2004 *in press a*). The impact of memory demands on audience design during language production. *Cognition*.

Horton, W.S., & Gerrig, R. J. (2004 *in press b*). Conversational common ground and memory processes in language production. *Discourse Processes*.

Horton W.S. & Keyser B. (1996) When do speaker take common ground? *Cognition, 59,* 91- 117

Keysar, B, Lin, S., Barr, D. J, (2003). Limits on theory of mind use in adults. *Cognition*, *89*, 25–41

Lockridge., C. B, & Brennan, S. E (2002). Addressees' needs influence speakers' early syntactic choices *Psych Bull & & Rev*, 9, 550-557

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects in the comprehension of referring expressions. *JML, 49*, 201–213

Pickering, M., & Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behav Brain Sci, 27*, 169-190.

Pomplun, M., Rieser, H., Ritter, H. & Velichkovsky, B. M. (1997). Augenbewegungen als kognitionswissenschaftlicher Forschungsgegenstand. In Kluwe, R.H. (Ed.), *Kognitionswissenschaft: Strukturen und Prozesse intelligenter Systeme*, 65-106. Wiesbaden: Deutscher Universitätsver.