Referential Form, Word Duration, and Modeling the Listener in Spoken Dialogue

E. G. Bard (ellen@ling.ed.ac.uk)
HCRC and Department of Theoretical and Applied Linguistics, Adam Ferguson Building, University of Edinburgh
Edinburgh EH8 9LL, U.K

M. P. Aylett (matthewa@cstr.ed.ac.uk)
HCRC and Rhetorical Systems, 2 Buccleuch Place, University of Edinburgh,
Edinburgh EH8 9LW, U.K

Abstract

Referring expressions are thought to be tailored to the needs of the listener, even when those needs might be costly to assess, but tests of this claim seldom manipulate listener's and speaker's knowledge independently. The design of the HCRC Map Task enables us to do so. We examine two 'tailoring' changes in repeated mentions of landmark names: faster articulation and simplified referring expressions. Articulation results replicate Bard et al. (2000), depending only on what the speaker has heard. Change between mentions was no greater when it could be inferred that the listener could see the named item (Expt 1), and no less when the listener explicitly denied ability to do so (Expt 2). Word duration fell for speaker-Given listener-New items (Expt 3). Reduction was unaffected by the repeater's ability to see the mentioned landmark (Expt 4). In contrast, referential form was more sensitive to both listener- (Expt 3) and speaker-knowledge (Expt 4). The results conform most closely to a Dual Process model: fast, automatic, processes let the speaker-knowledge prime word articulation, while costly assessments of listener-knowledge influence only referential form.

Introduction

Speakers are said to design their utterances to suit the needs of their listeners, insofar as those needs can be known (Ariel, 1990; Clark & Marshall, 1981; Gundel, Hedberg, & Zacharski, 1993; Lindblom, 1990). Certainly, there is variation in form. Clarity of pronunciation varies with predictability from local context (Hunnicutt, 1985; Lieberman, 1963) and with repeated mention (Fowler & Housum, 1987). Referential forms are syntactically simpler the more readily interpreted or 'accessible' their antecedents, are (*a blacksmith's cottage* v *it*) (Ariel, 1990, Fowler, Levy, & Brown, 1997; Gundel, et al., 1993; Vonk, Hustinx, & Simmons, 1992). Yet maintaining an incrementally updated model of what the listener knows, what is established common ground, and what the listener needs to know is a considerable cognitive task. Because speaker's and listener's knowledge overlap and because it may be impossible to assess the latter accurately, speakers may default to an account of their own knowledge as a proxy for the listener's (Clark & Marshall, 1981). In fact, many studies simply assume that the two are the same: they manipulate the speaker's knowledge without independently manipulating the listener's (see Keysar, 1997).

This paper compares two versions of the hypothesis that referring expressions are genuinely tailored to the addressee. One deals with the articulation of individual words, the other with the syntactic form of referring expressions. Under current models of language production, NP structure and articulation are generated within units of different sizes, intonational or syntactic phrases on the one hand and phonological words, lexical words, or syllables on the other (Levelt & Wheeldon, 1994; Smith & Wheeldon, 1999; Wheeldon & Lahiri, 1997). Moreover, speech appears to be produced in a cascade, with a sequence of smaller units being prepared for articulation even as the succeeding larger unit is being designed. Thus, incrementally updating a listener model in order to articulate each phonological word appropriately would impose a much heavier computational burden than updating it phrase by phrase. Making both kinds of update for the processes running in parallel would be even more demanding, with the listener model operating both in the state appropriate to the most recently produced word and in the state created by the most recently planned phrase.

We will first develop existing hypotheses about how speakers model listeners while planning and producing speech. Then we will report four studies which test these hypotheses on materials from a single corpus. They follow the comparisons made by Bard et al. (2000) on a psychological measure of clarity, the intelligibility to naïve listeners (recognition rate) of a balanced sample of excised spoken words. The present paper reports a phonetic measure of clarity (word duration), and a syntactic measure of referential form for all suitable cases in a dialogue corpus. Finally, we will discuss the implications of the comparison.

Modeling Listeners while Speaking

Existing accounts of tailoring to listeners' needs make different computational demands on speakers. Where they are not designed with a view to on-line processing, we will attempt to interpret their implications.

Lindblom's *H-and-H Hypothesis* (1990) makes the heaviest computatonal demands. It posits that speakers adjust the articulation of spoken words to the knowledge which the listener can currently recruit to decoding the speech signal: speakers hyper-articulate when listeners lack such auxiliary information and hypo-articulate when redundancy is high. More redundant linguistic environments do contain word tokens articulated with greater speed and less precision (Bard & Anderson, 1983, 1994; Fowler & Housum, 1987; Hunnicutt, 1985; Lieberman, 1963; Samuel & Troicki, 1998). The question is whether this relationship depends on the speaker's consulting an up-to-date model of the listener's current knowledge each time s/he prepares the prosodic character of a phonological word or the articulation of its syllables. Though H-and-H does not preclude defaulting to speaker-knowledge, it is framed in terms of genuine listener-knowledge and implies that speakers should observe listeners continuously for signs of misunderstanding or disagreement. Wherever speaker's and listener's knowledge differ, the latter should take precedence.

In contrast, Brown and Dell (1987) propose a modular division between the initial formulation of utterances and the revision of output which does not adequately convey the intended concepts. The listener's knowledge is implicated only in revision (Dell & Brown, 1991, pp. 119-120). Called the *Monitor and Adjust Hypothesis* (Horton & Keysar, 1996), this model defaults to speaker-knowledge first and pays later – if necessary. As originally formulated, Monitor and Adjust does not explain how the hitherto speaker-driven processes assess the adequacy of an utterance from the listener's point of view. We assume that each interlocutor's knowledge includes a record of what the other has actually said. Listeners' occasional explicit feedback, a minimal listener model, could therefore influence a modular system which revises inadequate utterances. Under this Extended Monitor and Adjust Hypothesis, post-feedback utterances could reflect any listener-knowledge which the feedback has conveyed. Otherwise, listener-knowledge should be irrelevant to production.

The third proposal deals with co-presence, characteristics of listeners which affect likely overlap with speakers' own knowledge (Brennan & Clark, 1996; Fussell & Krauss, 1992; Isaacs & Clark, 1987; Schober, 1993). The manifestations of co-presence in the dialogue literature are many, but the notion was originally used to reduce the computation which a speaker must perform to determine the unknown component of mutual knowledge, i.e. what the listener knows. Under this heading, Clark & Marshall (1981) list shared community membership, physical co-presence of interlocutors and the objects under discussion, and knowledge both of the dialogue and of a suitable scenario. Since much of co-presence is long-lasting, it can reduce both the depth and the frequency of listener modeling. To exploit these economies, speakers should attend to evidence for and against co-presence, and they should maintain defaults for some undefined time after positive evidence. We will call this the *Co-presence Default Hypothesis.*

Finally, Bard et al. (2000) develop a suggestion of Brown and Dell (1987) which we will call the *Dual Process Hypothesis*. It proposes a division between fast, automatic processes, which have no computational cost, and slower, more costly processes requiring inference or attention. The former include priming. (Balota, Boland, & Shields, 1989; Mitchell & Brown, 1988), an effect of the speaker's own recent experience. The latter include the kind of complex reasoning usually implicated in constructing a model of the listener. In competition with listener-modelling are the computations which support planning a dialogue or tracking a shared task. When there is competition for time and attention, (Horton & Keysar, 1996), the inferential processes may suffer, leaving the speaker with only cost-free defaults in the form of his or her own knowledge.

Studies of Intelligibility and Referring Expression

Givenness and Referring Expressions

To test these hypotheses, we made use of two effects of Given status broadly defined. First, spoken words introducing New items are longer and clearer than those in repeated mentions (Fowler & Housum, 1987) but only when the two tokens are co-referential (Fowler, 1988; Bard et al, 1991). Initial mentions of items uttered without visible referents (Prince's (1981) 'brand new') are also longer and clearer than those with visible referents (Prince's (1981) 'situationally' Given) (Bard & Anderson, 1994). Second, referring expressions simplify with repeated mention (*a blacksmith's cottage.... it)* as their antecedents become more accessible (Ariel, 1990, Gundel, Hedberg, & Zacharski, 1993). To compare the two systems, we used a single coded corpus of spontaneous speech which made it possible to select items which were Given to one or both interlocutors on the basis of what each saw, said, or heard in the dialogue.

Method

Materials. Materials came from the HCRC Map Task Corpus (Anderson et al., 1991), 128 unscripted dialogues in which pairs of Glasgow University undergraduates ($N = 64$) communicated routes defined by labeled cartoon landmarks on schematic maps of imaginary locations. Instruction Giver's and Follower's maps for any dialogue matched only in alternate landmarks. Participants knew that their maps might differ but not where or how. Players could not see each other's maps. Familiarity of participants and ability to see the interlocutor's face were counterbalanced. Each participant served as Instruction Giver for the same route to two different Followers and as Instruction Follower for two different routes.

Channel per speaker digital recordings were word-segmented. All words of any expression referring to a landmark were coded for the landmark, tagged for part-of-speech, and parsed. Interrupted or disfluent items were excluded. All remaining expressions making repeated reference to a landmark and meeting experiments' design criteria were used. Duration was measured only if both mentions include the same words. All repeated mentions were are assessed for syntactic form.

Dependent Variables

K-reduction. Normalized duration (Campbell & Isard, 1991) assigns each word token a value, k, representing its position in the expected log length distribution for words of its dictionary phoneme composition and stress pattern. *K-reduction* is the difference between the k-durations of a read control form and of the corresponding item in running speech. Faster articulation with repeated mention would enhance k-reduction.

Form of referring expression. The 27 items with relative clauses in their first mentions were excluded because of a conflict in coding schemes. All other first and second mentions of landmarks ($N = 1136$) were classed on the scale displayed in Table 1, where '0' indicates least simplified/accessible. Simplification score should increase with repeated mention.

Table 1. Simplification scale for referring expressions

Code	Definition		Example
0	numeral +⌉ indef art +⌋	noun sequence	*one mountain* *a mountain*
1	def art +⌉ poss +⌋	nominal	*the mountain* *my one*
2	deictic adj+	possess pro deictic pro nominal	*mine* *that* *this mountain*
3		other pro	*it*

Experiment 1: Inferable Listener Knowledge

Design. Experiment 1 compared repeated mentions of landmarks appearing on both players' maps in two conditions, self- and other-repetition. The key to the design is the fact that a speaker who first mentions a landmark must have it on his or her own map. Thus, in an other-repetition the repeater can easily infer that the introducer can see the landmark. The second token, therefore, refers to an object which is Given both to the repeater who has heard it mentioned and can see it, and to the current listener who has also heard it mentioned, who can see it, and who has mentioned it. Self-repetitions differ in two respects: the repeater who introduced the landmark does not know if the listener can see it. Thus, the design contrasts a case where the listener can easily be concluded to have more knowledge of the referent with one where the listener's knowledge is in doubt. The inference about shared visual resource is both simple and important to the task. Since visibility can affect clarity of mention (Bard & Anderson, 1994), tailoring to the listener here should enhance change across mentions (more k-reduction, greater simplification of expression) where the listener has more information - in other-repetition.

Not all hypotheses make this prediction. H-and-H predicts that articulation will be sensitive to the listener's needs in this way. Dual Process predicts instead that any effect will be found in referential form, which is designed over intervals long enough to allow for completing the necessary inference. Copresence predicts effects to what the listener can see. Monitor and Adjust makes no special prediction because speakers are not obliged to model listeners cotinuously to conduct dialogues.

Table 2. Changes with self- v other-repetition:

Measure	Original speaker	
	Self	Other
Articulation: k-reduction	0.127	0.192
Form of referring expression	0.878	0.745

Results. Table 2 shows similar changes in articulatory clarity for self- and other-repetition. As in Bard et al. (2000), words were said faster on repeated mention ($F_2 (1,691) = 63.75$, $p < .0001$) but with no significant difference in reduction between the 263 other-repetitions and the 430 self-repetitions (mention x prior speaker: n.s.). Form of referring expression simplified with repeated mention ($F_2 (1,269) = 177.12$, $p < .0001$) but again did not distinguish the 90 other-repetitions from the 430 self-repetitions (mention x prior speaker:

n.s.). Contrary to the H-and-H predictions, the listener's experience was not critical. Repetitions of any mentions of visible objects which the repeater had heard were treated alike.

Experiment 2: Listener Feedback

Design. Experiment 2 provides a more direct test of the effects of listener knowledge. When one speaker introduces an unshared landmark, the listener, who lacks it, may provide corrective feedback indicating the discrepancy between the players' maps. Sometimes, however, the listener fails to do this. We compare repeated mentions of the names of unshared landmarks by the same speaker in these two cases. In both, the repeater has said and heard the initial mention and can see the object. When the listener denies having it, the repeater knows that the listener has heard the word but cannot see the object. Otherwise, the repeater cannot tell if s/he can see the landmark.

Cooperative behavior would yield a more restricted effect of repetition where the listener has denied ability to find the object. This comparison is important for the Extended version of Monitor and Adjust, which predicts that feedback at least could make a difference to subsequent mention design. Only Dual Process holds that pronunciation must and syntactic form may be designed without regard to the listener's comments.

Table 3. Changes for repetition with v without feedback on listener's inability to see the landmark:

Measure	Visibility to listener	
	Not denied	Denied
Articulation: k-reduction	0.070	0.140
Form of referring expression	0.470	0.410

Results. Table 3 shows that both articulation and form of referring expression were unaffected by feedback. The 73 repetitions with intervening denial and the 122 without abbreviated with repetition significantly and equally ($F_2(1,193) = 9.45$, $p = .0024$; mention x denial: *n.s.*). The simplification of referring expressions on second mention was similar for the 44 cases with intervening denials and the 86 without ($F_2 (1,128) = 18.49$, $p < .0001$; mention x denial: *n.s.*). Feedback that could block defaulting under Monitor and Adjust does not do so. Only what the repeater has seen, heard, and said seems to play a role.

Experiment 3: Listener Identity

Design. Experiment 3 examines introductory mentions of the same shared landmarks in Givers' two trials with the same map. In the first trial, the landmark is New for both players. In the second, it is Given for the speaker, an Instruction Giver who has mentioned it before, heard that mention, and seen the landmark. However, it is New to each successive listener. Adjustment to the new listener should block any tendency to utter a second introduction as a shorter, more accessible Given item.

In fact, this experiment offers the classic test of Co-presence. Mentions to new listeners should be geared to their ignorance. H-and-H posits that listener modeling will block at least articulatory change. Monitor and Adjust predicts changes in articulation and form, because the speaker's knowledge controls language production, not the listener's. Dual Process predicts a loss of clarity because articulation depends on the speaker's previous mention, not on the listener's knowledge. Only form of referring expression may reflect the listener's ignorance and remain unchanged.

Results. Table 4 shows that second introductions are significantly shorter than first for 239 pairs of words (F_2 (1,238) = 12.48; $p < .0005$). In contrast, simplification of referring expression does not significantly increase over 116 pairs of introductory mentions (F_2 (1,115) < 1). Thus, word reduction appears to reflect the Given status of the item for the speaker, while referential form reflects the fact that the freshly introduced landmark is New for each listener. Greater sensitivity in form of referring expression is predicted only by the Dual Process Hypothesis.

Table 4. Change with reintroductions to new listeners.

Measure	Introduction	
	1	2
Articulation: k-reduction	0.498	0.558
Form of referring expression	0.466	0.552

Experiment 4: Speaker Knowledge

Design. In Experiment 4, only other-repetitions were used, but now the landmark in question was either shared by both speakers or absent from the repeater's map. In both cases, the original introducer, who is the listener at the point of second mention, can see the item, has mentioned it, and has heard it mentioned. The repeater has also heard it mentioned, but has not mentioned it and may or may not be able to see it.

Because this experiment holds listener knowledge constant, adjustment to the listener cannot yield any differences between conditions. If the speaker's visual surroundings are important, then changes across

repeated mentions will be greater for visible, shared landmarks than for unshared.

H-and-H predicts no effect of what the speaker can see. Monitor and Adjust allows the speakers' knowledge to affect both dependent variables. Dual Process claims that auditory priming keys articulation to speaker-knowledge, but also allows for costly access to additional information, which would permit effects of speaker-knowledge on referential form.

Table 5. Changes on other-repetition of shared v unshared landmark names

Measure	Visibility to speaker	
	+	–
Articulation: k-reduction	0.114	0.183
Form of referring expression	0.745	0.240

Results. Table 5 shows reduction of word tokens with repeated mention (F_2 (1,224) = 12.37, $p < .0005$) but .no significant difference between the outcome for the 144 shared, visible landmarks and the 82 unshared (mention x visibility: *n.s.*). Referential form, however, is speaker-centric. Second mentions are more simplified than first overall (F_2 (1,138) = 24.67, $p < .0001$), but the change is greater for the 90 shared items than for the 50 unshared (mention x visibility: F_2 (1,138) = 6.48, $p < .02$).

This outcome is not consistent with adjustment to listeners alone or with overall use of speakers'-knowledge as a proxy for listeners'. It conforms best to the notion that referential form is sensitive to a wider range of information than articulatory clarity.

Discussion

The experiments reported here and in Bard et al. (2000) test for effects on repeated mentions of several aspects of speaker- or listener-knowledge. Experiment 1 pitted the speaker's own experience in seeing and hearing against the listener's under two conditions, when it could and could not readily be inferred that those listeners could see the landmark. Experiment 2 pitted the speaker's experience of seeing, saying, and hearing against the listener's declared inability to see the item in question. Experiment 3 pitted the speaker's experience in having seen the mentioned landmark, mentioned it, and heard it mentioned against the new listener's ignorance of the item as the landmark was introduced in a second trial. Experiment 4 kept the listener's knowledge constant as well as the speaker's experience in hearing a prior mention, but manipulated the speaker's ability to see the landmark.

In all these cases, the repeating speaker had heard the original mention. In all, clarity of articulation was

sensitive only to what the speaker had heard. These are exactly the results found by Bard et al. (2000) for a balanced but restricted sample of materials and with intelligibility to naïve listeners directly measuring clarity. Thus, reduction and consequent reduction in articulatory detail with repeated mention is conditioned by the repeater's experience. There is no indication that models of the listener are consulted.

Referential form showed a different pattern. Like articulation, it was insensitive to some information which should have entered a model of the listener: an indications of what the listener could or could not see (Expts 1 and 2). Yet, it did show two effects which articulation did not. First, referential form did not simplify on re-introduction to new listeners (Expt 3)[1]. In this case, form of referring expression was tailored to the listener's needs. Second, simplification of form across repeated mentions was enhanced when the speaker could see the named landmark (Expt 4). Thus, referential form is more sensitive than articulation but to both interlocutors' knowledge.

Why should form have these characteristics? Form of referring expression does not respond on-line to aspects of co-presence delivered via feedback or inference. It shows neither the complete insensitivity to listeners that Monitor and Adjust predicts for initial design, nor the sensitivity to feedback which should guide redesign. We would argue that Map Task participants juggled competing demands on their attention, as the Dual Process Hypothesis predicts. Unlike the fast automatic processes through which speaker memory affects articulation, slower processes can compete for attention with the communicative task in hand. In this task, however, listener modeling does not take precedence. Only the listener factor most grossly related to the task -- who is participating – affects the design of referring expressions. It shares that honor with an equally basic speaker factor -- what is on the speaker's own map.

The difficulty of the communicative task may well influence the degree to which speakers appear to be modeling their listeners. Certainly direct manipulation of communicative tasks changes speakers' priorities (see Horton & Keysar, 1996). Presumably, speakers could be more sensitive to listener-knowledge if some kind of external record-keeping were to ease the computational burden. The Dual Process Hypothesis predicts that both task and memory load should influence the design of referring expressions, but that neither should affect the articulation of individual words.

---

[1] While Jurafsky et al. (2001) have recently reported less reduction for reintroduction to new listeners than to old, they find significant reduction to both.

References

Anderson, A. Bader, M., Bard, E.G., Boyle, E., Doherty, G., et al. (1991). The H.C.R.C. Map Task Corpus. *Language and Speech*, *34,* 351-366.

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents.* London: Routledge/Croom Helm.

Balota, D. A., Boland, J. E., & Shields, L. W. (1989). Priming in pronunciation: Beyond pattern-recognition and onset latency. *Journal of Memory and Language*, *28,* 14-36.

Bard, E. G., & Anderson, A. (1983). The unintelligibility of speech to children. *Journal of Child Language*, *10,* 265-292.

Bard, E. G., & Anderson, A. (1994). The unintelligibility of speech to children: Effects of referent availability. *Journal of Child Langua*ge, *21,* 623-648.

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

Brennan, S., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning Memory and Cognition, 22,* 1482-1493.

Brown, P., & Dell, G. (1987). Adapting production to comprehension -- the explicit mention of instruments. *Cognitive Psychology, 19,* 441-472.

Campbell, W. N., & Isard, S.D. (1991). Segment durations in a syllable frame. *Journal of Phonetics, 19,* 37-47.

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.* Cambridge: Cambridge University Press.

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.* Hillsdale: Erlbaum.

Fowler, C. (1988). Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech, 28,* 47–56.

Fowler, C., & Housum, J. (1987). Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, *26,* 489-504.

Fowler, C., Levy, E. & Brown, J. (1997). Reductions of spoken words in certain discourse contexts. *Journal of Memory and Language*, *37,* 24-40.

Fussell, S., & Krauss, R. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology, 62,* 378-391.

Gundel, J.K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69,* 274-307.

Horton, W., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59,* 91-117.

Hunnicutt, S. (1985). Intelligibility vs. redundancy — conditions of dependency. *Language and Speech*, *28,* 47-56.

Isaacs, E., & Clark, H. H. (1987). References in conversation between experts and novices. *JEP: General, 116,* 26-37.

Jurafsky, D., Bell, A., Gregory, M., Raymond, W., & Girand, C., (2001). Probabilities in the mental grammar: evidence from language production in natural conversation. *Proceedings of CUNY2001*, Philadelphia, PA..

Levelt, W., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition, 50,* 239-269.

Lieberman, P. (1963). Some effects of the semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6, 172–175.

Lindblom, B. (1990). Explaining variation: a sketch of the H and H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling.* Dordrecht, Netherlands: Kluwer Academic Publishers.

Keysar, B. (1997). Unconfounding common ground. *Discourse Processes, 24,* 253-270.

Mitchell, D. B., & Brown, A. S. (1988). Persistent repetition priming in picture naming and its dissociation from recognition memory. *JEP:LMC, 14,* 213-222

Prince, E. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical Pragmatics.* New York: Academic Press.

Samuel, A., & Troicki, M. (1998). Articulation quality is inversely related to redundancy when children or adults have verbal control. *Journal of Memory and Language*, *39,* 175-194.

Schober, M. (1993). Spatial perspective taking n conversation. *Cognition, 47,* 1-24.

Smith, M, & Wheeldon, L (1999). High level processing scope in spoken sentence production. *Cognition*, *73,* 205-246.

Vonk, W., Hustinx, L., & Simmons, W. (1992). The use of referential expressions in structuring discourse. *Language and Cognitive Processes, 7,* 301-33.

Wheeldon, L., & Lahiri, A. (1997). Prosodic units in speech production. *Journal of Memory and Language*, *37,* 356-81