Replicability of Transaction and Action Coding in the Map Task Corpus*

Amy Isard and Jean Carletta
Human Communication Research Centre
University of Edinburgh
Email: J.Carletta@edinburgh.ac.uk

Abstract

Task-oriented dialogues can normally be divided into subdialogues, each of which reflects collaboration on a particular substep of the task, and which we call 'transactions'. We have devised a way of identifying transactions and their associated actions for HCRC Map Task dialogues, and we have tested the replicability of our coding scheme using naive subjects.

Introduction

Much work on dialogue has concentrated on dialogues arising from collaborative tasks. These dialogues are both easier to analyse than free-form conversations and more relevant to practical applications of human-computer dialogue. At least from the participants' points of view, the most important issues are how to break the task into executable subtasks and what actions to perform and when. Comparatively little work has been done which relates participants' actions to what happens in the dialogue. During task-oriented dialogue, the participants form collaborative plans to reach a joint goal, transferring information back and forth to each other both about how to do the steps in the task and about the current state of the plan. Often they work through the task in an orderly fashion, performing actions as they go, and the dialogue can be divided into subsections which reflect the steps of the task, similar to Sinclair and Coulthard's *transactions* (Sinclair & Coulthard 1975). We have devised a way of identifying transactions and their associated actions for HCRC Map Task dialogues, and we have tested the replicability of our coding scheme on naive subjects.

*This work was supported by an Interdisciplinary Research Centre Grant from the Economic and Social Research Council (U.K.) to the Universities of Edinburgh and Glasgow.

The Map Task

The HCRC Map Task Corpus (Anderson *et al.* 1991) is a collection of 128 task-oriented dialogues involving approximately fifteen hours of speech. In the dialogues, two participants (who are either familiar or unfamiliar with each other and are separated by either a full-height or a half-height partition) have slightly different versions of a simple map with approximately fifteen landmarks on it. The maps may have different landmarks or have some of the landmarks in different locations. In addition, one participant has a route drawn on the map. The task is for the second participant to duplicate the route. An example route giver map is given in figure 1. The trials vary over whether or not the participants know each other and whether or not the partition allows eye contact. No partition allows the participants to see each other's maps.

The Map Task is an ideal first domain for studying the relationship between actions and dialogue structure because there is an obvious basic order to the route which route givers nearly invariably follow, even though they can choose to chunk the route in different ways. In addition, there are only two possible route follower actions (drawing a route segment and crossing a previously drawn segment out), both of which are captured on video (at least for part of the corpus).

The Coding System

In most map task dialogues, the route giver breaks the route into manageable segments and describes each of them one by one. Our coding system has two components: we code both (1) how route givers divide conveying the route into subtasks and what parts of the dialogue serve each of the subtasks, and (2) what actions the route follower takes and when.

Figure 1: An Example Route Giver Map

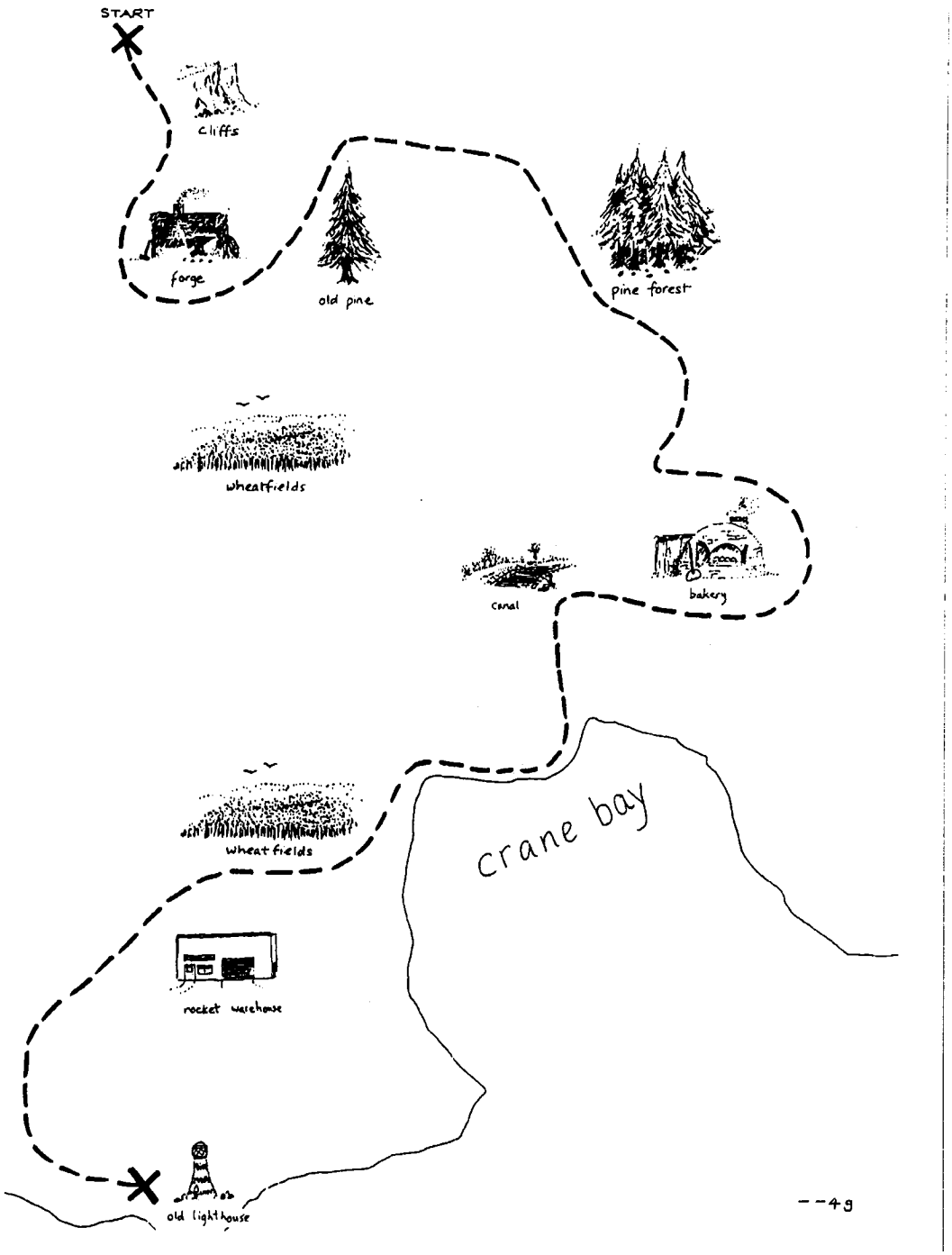

Figure 2: An Example Overview Transaction

G: And what we're basically going to be... where we're basically going to be going is towards.. I'll t-say this sort of globally
F: Mmm-hmmm.
G: then I'll do it more precisely. What we're basically doing is going, erm, south-east and then, erm, north-east, so you can imagine... a bit like a diamond shape if you like.
F: Mmm.
G: Southeast then northeast
F: Mmm.
G: and then northwest and then north, but the line's a lot more wavy than that. I'm just trying to give you some kind of overall picture.
F: Mmm.
G: It may not be very useful but.

Our basic route giver coding identifies the start and end of each segment and the subdialogue which conveys that route segment. However, map task participants do not always proceed along the route in an orderly fashion; as confusions arise, they often have to return to part of the route which were previously discussed and which at least one of the participants thought had been successfully completed. In addition, participants occasionally overview an upcoming segment in order to provide a basic context for their partners, without the expectation that their partners will be able to act upon their descriptions, as in the transaction in figure 2. They also sometimes engage in subdialogues which are not relevant to any segment of the route, sometimes about the experimental setup but often nothing at all to do with the task. This gives us four transaction types: 'normal', 'review', 'overview', and 'irrelevant'. Coding involves marking in the dialogue transcripts where a transaction starts and which of the four types it is, and for all but 'irrelevant' transactions, indicating the start and end point of the relevant route section using numbered crosses on a copy of the route giver's map. We do not explicitly code the endings of transactions because, generally speaking, transactions are large enough that they do not appear to nest; if a transaction is interrupted to, for instance, review a previous route segment, by and large participants restart the goal of the interrupted transaction afterwards rather than picking up where they left off. Note that it is possible for several transactions (even of the same type) to have the same starting point on the route.

Our basic route follower coding identifies whether the follower action was drawing a segment of the route or crossing out a previously drawn segment, the start and end points of the relevant segment, indexed using numbered crosses on a copy of the route follower's map.

The Replication Study

Since we can tell what the route follower drew and when they drew it from the video, we expect this coding to be uncontroversial. However, determining the transaction structure of a dialogue and choosing points which reflect the route segment which the route giver was trying to convey during a transaction requires judgment. In order to test whether or not our coding system for this information is replicable, we taught four naive subjects how to use it and had them all code the same four dialogues so that we could compare their coding to each other and to that of the first author. All four subjects were postgraduate students at the University of Edinburgh studying Cognitive Science, Artificial Intelligence or Maths; none of them had prior experience of the Map Task or of dialogue or discourse analysis. The four dialogues (eaq4c1-4 from the Map Task Corpus) were all from the eye contact condition but evenly mixed the other corpus conditions. All four dialogues used different maps and differently shaped routes.

We judged that the task would be too difficult if naive subjects were required to read the transcripts, listen to the tapes, and look at the maps all at the same time, so we did not give subjects access to the spoken dialogues. We prepared the dialogue transcripts so that they could be easily read, labelling each turn as coming from the giver or the follower and preserving the transcript conventions for marking overlapped speech but removing the microtagging for non-words. Conversational moves in the framework of Kowtko et al. (Kowtko, Isard, & Doherty 1992) divide dialogues into utterances which serve different goals in the collaboration; for instance, 'instruct' moves instruct the partner to take some action, 'query-yn' moves ask yes-no questions, and 'acknowledgement' moves show acceptance of some information has been given. A special case is 'ready' moves, which are discourse markers signaling that the speaker is moving on to a new goal. Intonational cues are necessary for deciding whether instances of phrases such as 'OK' and 'right' are ready or acknowledgement moves. We did not want the subjects to be allowed to mark transaction boundaries just anywhere because they would not have been able to tell on the basis of the written transcripts what side of these phrases to place the boundary on. However, transaction boundaries can reasonably be expected to occur at move boundaries, since transactions are defined in terms of larger participant goals. Therefore we

put blank lines in the transcript at each move boundary, but omitting the boundaries after ready moves. Subjects were allowed to place transaction boundaries at any of the blank lines. The move codings themselves were not present on the transcripts.

Each subject was given a set of instructions with short examples and a sample dialogue extract and pair of maps to take away and read at leisure; they were asked to return with the dialogue extract coded. When they returned they were given a chance to ask questions to clarify any part of the instructions which they had not understood, and when everything had been explained to their satisfaction, they were given the four complete dialogues and maps to take away and code in their own time. The four subjects did not speak to each other about the exercise. Three of the four subjects asked for clarification of the 'overview' distinction; there were no other queries.

A note about replicability statistics

Previous work using annotative codings for discourse and dialogue has not agreed on how to show replicability of results. For instance, Kowtko et al. (Kowtko, Isard, & Doherty 1992), having devised a coding scheme for conversational moves involving thirteen distinctions, cite separate pairwise agreement percentages between one expert coder and each of three naive coders in order to argue that the coding scheme is replicable. Meanwhile, Passonneau and Litman (Passonneau & Litman 1993), in arguing that naive subjects can reliably agree on whether or not given prosodic phrase boundaries are also discourse segment boundaries, measure replicability using 'percent agreement', defined (in (Gale, Church, & Yarowsky 1992)) as the ratio of observed agreements with the majority opinion to possible agreements with the majority opinion. Neither of these approaches successfully conveys whether or not a coding scheme is replicable, most importantly because they both fail to take into account the level of agreement one would expect random coders to attain.

Chance expected agreement depends on the number and relative proportions of the categories used by the coders. For instance, suppose that two coders place objects in one of two categories in equal proportions, but choose the category randomly. We would expect them to agree with each other half of time. If instead, the two coders were to use four categories in equal proportions, we would expect them to agree 25% of the time (since no matter what the first coder chooses, there is a 25% chance that the second coder will agree.) And if both coders were to use one of two categories, but use one of the categories 95% of the time, we would expect them to agree 90.5% of the time ($.95^2 + .05^2$). This makes it impossible to interpret raw agreement figures. It is possible to work out expected agreement for the statistic used by Passonneau and Litman but it is somewhat more difficult because the statistic itself guarantees at least 50% agreement by only pairing off coders against the majority opinion. Passonneau and Litman note that their figures are not properly interpretable and attempt to overcome this failing to some extent by showing that the agreement which they have obtained at least significantly differs from random agreement. However, all this tells us is that it is safe to assume that their coders were not coding randomly — reassuring, but no guarantee of replicability.

Conveying the scale of agreement is inherently problematic in nominally scaled data, and no measure will allow for fully comparable results. However, our concerns are largely the same as those in the field of content analysis (see especially (Krippendorff 1980) and (Weber 1985)), which has been through the same problems as we are currently facing and in which strong arguments have been made for using the kappa coefficient of agreement (Siegel & N. J. Castellan 1988) as a measure of replicability.[1]

The kappa coefficient (K) measures pairwise agreement among a set of coders placing things into categories, correcting for expected chance agreement.

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of times that the coders agree and $P(E)$ is the proportion of times that we would expect them to agree by chance, calculated along the lines of the intuitive argument presented above. (For complete instructions on how to calculate K, see (Siegel & N. J. Castellan 1988).) When there is no agreement other than that which would be expected by chance K is zero. When there is total agreement, K is one. When it is useful to do so, it is possible to test whether or not K is significantly different from chance (all of the results we cite are), but more importantly, loose interpretation of the scale of agreement is possible. Krippendorff (Krippendorff 1980) discusses what makes an acceptable level of agreement, while giving the caveat that it depends entirely on what one intends to do with the coding. For instance,

[1] There are several variants of the kappa coefficient in the literature, and in fact, Krippendorff's α and Siegel and Castellan's K differ slightly in the assumptions under which expected agreement is calculated. Here we use Siegel and Castellan's K, but α is so closely related in scale, especially under the usual expectations for replicability studies, that Krippendorff's statements about α hold, and we conflate the two under the more general name 'kappa'.

he claims that it is often impossible to find associations between two variables which both rely on coding schemes with $K < .7$, and says that content analysis researchers generally think of $K > .8$ as good replicability, with $.67 < K < .8$ allowing tentative conclusions to be drawn. In addition to his caveat, we would add that coding discourse and dialogue phenomena is inherently more difficult than many previous types of content analysis (for instance, coding occurrences of particular words in a document, or the number of times that someone uses his or her telephone). Whether this makes the method inappropriate for some discourse and dialogue work remains to be seen; our point here is merely that if as a community we adopt clearer statistics, that will illuminate both our individual results and the way forward.

There are several possible uses of the kappa coefficient within empirical work on discourse and dialogue, depending on whether or not one of the codings is considered definitive (the 'expert', or in some cases a coding done by a better but more expensive method). If there is no definitive coding, one might wish to describe the level of agreement amongst the coders as a group. If there is an expert coding, then it is possible to compare the group of naive coders to the expert coder both individually (in the style of Kowtko) and as a set (in the style of Passonneau and Litman, if one accepts their argument that the majority opinion actually reflects something). Comparing the naive coders to the expert as a set gives a general measure of replicability, and comparing them individually allows us to see if there are any 'odd men out' — some coders will simply do a better job than others, for whatever reason. In addition, even if there is an expert coding, it can be useful to look at agreement among the naive coders as a group; since often the expert coder is the one who wrote the coding instructions, if the naive coders agree among themselves better than they agree with the expert, it could be that the instructions mean something, but not what the expert intended! In this work we quote results for several different uses of the kappa coefficient.

Agreement on whether a move boundary is a transaction boundary

In our study, we have four naive coders and one expert classifying each move boundary as either being or not being a transaction boundary. We first look at agreement with the expert for each naive coder separately. There were 657 move boundaries for coders to mark; any one coder marked roughly one-tenth of them as transaction boundaries. Pairwise agreement, expected agreement, and kappa coefficients for each pairing of a naive coder with the expert are given in figure 3.

Figure 3: Agreement between the expert and naive coders individually

Coder	pairwise agreement	expected agreement	K
1	.94	.81	.68
2	.91	.81	.53
3	.94	.83	.65
4	.92	.86	.43

We next look at agreement among the coders as a group. Treating our expert the same as the other coders, pairwise agreement was 93%; in other words, counting over each possible pair of coders for each of the 657 move boundaries, coders agreed on whether the move boundary was a transaction boundary 93% of the time. We would expect 83% agreement by chance from coders marking the same proportions of move boundaries as transaction boundaries but deciding which ones were transaction boundaries randomly. This gives a kappa coefficient of .59.

The previous measure checked the amount of agreement among all coders. The next question to test whether or not our instructions conveyed what the expert coder intended by asking whether our naive coders agreed with each other more than they did with our expert coder. The results are exactly the same; 93% pairwise agreement with 83% expected by chance, with a kappa coefficient of .59. Therefore we conclude that any disagreement is general rather than reflecting a difference between expert and novice coders.

Agreement on where to place points on maps

All transaction boundaries except 'irrelevant' ones have points associated with them which give the beginning and ending of the route segment under discussion. Where the coders agreed that a boundary existed, they almost invariably placed the begin and end points of their segments within the same 1.5 inch segment of the route, as measured on the original A3 (296 x 420 mm) maps. Many of the points were closer together than that. In contrast, the closest points which did not refer to the same begin or end point of a transaction were almost invariably more than two inches apart and often much further. Therefore we are confident that these points really are intended to be the same, but we can not rely on placing these points more precisely than roughly to the nearest two inches.

Agreement on subcodes

We believe there to be generally good agreement on which type of transaction to mark at transaction boundaries, but given the small size of this study and the fact that the vast majority of transactions are of the 'normal' type, we can not show agreement figures. There were fifteen transaction boundaries agreed by all coders; one coder marked one of them as an 'overview', but all remaining agreed boundaries were marked as 'normal' by all coders. There were 78 boundaries marked by at least two coders; of these, 64 had no dissension on the sort of boundary which it was for the coders who marked it.

Sorting out disagreement

We have identified the following five areas of disagreement among our coders.

- Route givers often begin a route segment by considering possible descriptions and asking the route follower whether or not they have the landmarks involved. We intended these introductory questions to be included in the same transaction as the description which they facilitate. Sometimes coders placed the boundary after the introductory question, missing the connection with the following description. We suspect that if we were to specifically point out this situation, coders would be aware that they need to think carefully about it and they would code it more consistently. The vast majority of the disagreements of this type came from one naive coder in one dialogue, but instances of it occurred for all coders.

- Sometimes when they think a segment of the route has been finished, route followers ask questions such as "where do I go now?" We intended these questions to begin transactions, but sometimes the naive coders placed the question in the preceding transaction. The confusion was spread equally among the four naive coders but occurred only in one dialogue, probably reflecting that these questions are an idiosyncratic behaviour. Again, we suspect that if we were to specifically point out examples of this kind, coders would be able to mark these boundaries consistently. However, it is also possible that such questions will always be a problem area for this kind of coding, since they can simultaneously open a new transaction and close the previous one, if the route follower gives no other signal that the previous transaction's goal has been completed.

- The 'overview' coding proved to cause a lot of confusion. Subjects could not distinguish between the sequences overview-normal and normal-review where both transactions in the pair were about the same route segment, and also often included overview transactions in the normal transactions which follow. We believe that the confusion stems from the fact that true overview transactions are rare, so that coders try to apply the category to more prevalent behaviours. Therefore we have decided to omit this subcode in any future coding.

- Despite our placement of possible transaction boundaries to avoid confusion about discourse markers, there were still some 'right' and 'okay' moves which had possible boundaries either side, and subjects were not able to judge where the transaction boundary belonged. Correcting this confusion may be a matter of giving the subjects access to the speech and pointing out that there is a distinction, if that does not make coding so difficult that the coders make more errors overall. However, many possible uses of the coding would not require boundary placement to be this precise.

- One coder had many fewer transactions than the other coders, with each transaction covering a segment of the route which other coders split into two or more transactions, indicating that he thought the route givers were planning ahead much further than the other coders did. We believe this to be a standard problem encountered by anyone involved in discourse or dialogue segmentation to which there is no fully satisfactory solution. We believe that we could improve our own situation by having a longer training period with discussion between the expert and the novice, but any such discussion would inherently make it more difficult for others to apply the coding systems and hence would weaken its credibility.

Conclusions

We have described a method of coding transactions and actions for the map task corpus, and shown replicability results for this coding scheme using the kappa coefficient of agreement. Although K is used for replicability studies in other disciplines which use content analysis, it is not in common use within the discourse and dialogue communities. We have argued that it is the most appropriate metric in these circumstances. Although our current level of replicability might be acceptable for some purposes, this work has identified a number of confusions which it would be useful to correct at this stage. Our next step is to improve the coding scheme in an attempt to eliminate these confusions and increase its replicability.

References

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.; and Weinert, R. 1991. The HCRC Map Task Corpus. *Language and Speech* 34(4):351–366.

Gale, W.; Church, K. W.; and Yarowsky, D. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the Thirtieth Annual Meeting of the ACL.*

Kowtko, J. C.; Isard, S. D.; and Doherty, G. M. 1992. Conversational games within dialogue. Technical Report HCRC/RP-31, Human Communication Research Centre, University of Edinburgh.

Krippendorff, K. 1980. *Content Analysis: An introduction to its methodology.* Newbury Park, CA: Sage Publications.

Passonneau, R. J., and Litman, D. J. 1993. Intention-based segmentation: human reliability and correlation with linguistic cues. In *Proceedings of the 31st Annual Meeting of the ACL*, 148–155.

Siegel, S., and Castellan, N. John, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill, second edition.

Sinclair, J. M., and Coulthard, R. M. 1975. *Towards an Analysis of Discourse: The English used by teachers and pupils.* Oxford: Oxford University Press.

Weber, R. P. 1985. *Basic Content Analysis.* Newbury Park, CA: Sage Publications.