Finishing each other's ...
Responding to incomplete contributions in dialogue

Christine Howes, Patrick G. T. Healey, Matthew Purver, Arash Eshghi
{chrizba, ph, mpurver, arash}@eecs.qmul.ac.uk
Queen Mary University of London
Interaction, Media and Communication Research Group
School of Electronic Engineering and Computer Science, London E1 4NS, UK

Abstract

A distinguishing feature of dialogue is that contributions can be fragmentary or incomplete. Such incomplete utterances may be later completed by another interlocutor. These cross-person *compound contributions* (CCs) have been hypothesised to be more likely in predictable contexts but the contributions of different sources of predictability has not been systematically investigated. In this paper we present an experiment which artificially truncates genuine contributions in ongoing text-based dialogues, to investigate the effects of lexical, syntactic and pragmatic predictability of the truncation point on the likelihood of one's interlocutor supplying a continuation. We show that what is critical is the actual and presumed accessibility of common ground, and that while people are sensitive to syntactic predictability, this alone is insufficient to prompt a completion.

Keywords: Dialogue; compound contributions; common ground.

Introduction

It is well known that contributions to dialogue are often fragmentary or in some sense unfinished Fernández and Ginzburg (2002). These incomplete utterances may be subsequently completed, either by the original speaker following some response or interruption from an interlocutor, or, by another person (Purver et al., 2009).

These *compound contributions* (CCs) are a paradigmatic feature of dialogue, and cross-person CCs in particular are a key indicator of coordination between interlocutors. Although naturally occurring cross-person CCs and their interpretations have been studied (Lerner, 1996; Purver et al., 2009), there has not previously been a systematic, experimental, attempt to investigate the factors that influence how a completion for an incomplete utterance may be produced. Intuitively, people's willingness to finish another person's incomplete utterance will depend (at least) on how predictable the rest of the utterance is. There are several sources of possible predictability.

Expansions are CCs which add material (e.g. an adjunct) to an already complete syntactic element; *completions* are CCs which complete an incomplete element. Conversation analytic (CA) discussions of CCs suggest that they should preferably occur at *transition relevance places* (TRPs), points that are foreseeable by the participants. *Expansions* are CCs with split points at TRPs, and are more common in spoken dialogue (Howes et al., 2011) so ought to be more likely than completions.

Hypothesis 1 *Cross-person completions are more likely at transition relevance places*

Second, *completions* should tend to occur at syntactically projectable points (e.g. compound turn constructional units Lerner, 1991).

Hypothesis 2 *Cross-person completions are more likely when they are syntactically predictable.*

A third source of predictability comes from the degree to which the speaker and hearer share, or can be assumed to share, common ground relevant to the CC. If the topic of the utterance is already in the common ground then the content of the completion is more predictable.

Hypothesis 3 *Cross-person completions are more likely when they address topics that are part of the common ground.*

The effects of these different forms of predictability are directly tested here for the first time using a text chat experiment performed with the DiET experimental platform. The evidence points towards shared knowledge being a key factor with other sources of predictability also contributing.

Method

In this experiment, to see what factors influence how people respond to unfinished turns and their likelihood of producing a continuation, a number of genuine single contributions in dyadic text-based conversations were artificially split into two parts, using the DiET chat tool.

The DiET chat tool

The Dialogue Experimental Toolkit (DiET) chat tool is a text-based chat interface into which interventions can be introduced into a dialogue in real time. These interventions can take a number of forms; turns may not be relayed, additional turns may be added, as in Healey et al. (2003), in which spoof clarification requests are added to the dialogue, or turns may be altered prior to transmission. As these manipulations occur as the dialogue progresses, they cause a minimum of disruption to the 'flow' of the conversation.

The DiET chat tool is a custom built Java application, consisting of two main components: the server console and the user interface. The server time-stamps and

stores each key press, and acts as an intermediary between what participants type and what they see. All turns are passed to the server, from where it is relayed to the other participants. Prior to being relayed, real turns can therefore be automatically altered by the server or not relayed, or fake turns can be introduced.

Character-by-character interface In the character-by-character version of the DiET chat tool, the user interface consists of a single chat window. Below this, there is a status bar, which indicates if any participants are actively typing (see figure 1).

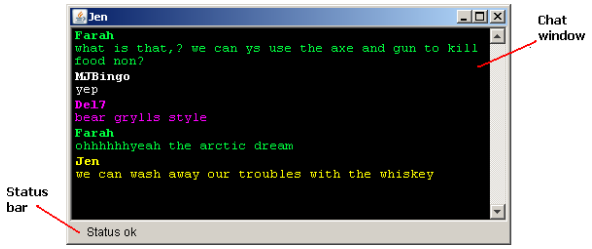

Figure 1: The DiET chat window (as viewed by *Jen*)

Unlike traditional chat interfaces (such as MSN Messenger), users type directly into the same window in which they see their interlocutors' contributions. This means that each character that any of the participants type is displayed in the window at the time it is entered – i.e. users see both their own and their interlocutors' contributions unfold in a character-by-character fashion. Consequently, only one participant may type at a time.

The intervention

For this experiment, single contributions were artificially split into two parts. The first part was transmitted to the other participant as it was typed, with the turn truncated according to various factors as discussed below. Following a pilot study, which showed that people were more likely to supply a response after a filler "..." or "...?" than if there were no filler (after a filler: 18/26, 69%, no filler: 12/45, 27%; $\chi^2_{(1)} = 12.24, p < 0.001$), the truncated first part of the genuine turn was followed by a text filler. Subsequently, there was a delay of 12 seconds, during which the other person could respond if they wished. Any response was trapped by the server and not relayed to the original sender, before the rest of the original (interrupted) contribution was transmitted.

Split points are manipulated according to measures of a) syntactic and b) lexical predictability calculated as each turn is produced.

Entropy

Entropy is a measure of uncertainty: the higher the entropy, the higher the uncertainty; and the lower the entropy, the higher the predictability. Here we used two measures: part-of-speech entropy, to capture the "syntactic" predictability of one part-of-speech (POS) following another; and lexical entropy, to capture the predictability of a particular lexical item following a specific POS. To illustrate the difference: although determiners are predictably followed by nouns, there are lots of different nouns: determiners therefore have a relatively low POS entropy, and a relatively high lexical entropy.

Since predictability depends on dialogue context and topic, entropy values were calculated from a corpus of prior dialogues (53663 word tokens) collected using the same tool and domain (the balloon task – see below).[1] POS tags were generated using the Stanford POS tagger (Toutanova et al., 2003) with a misspellings map for common chat abbreviations and typos. For each POS, entropy was calculated as follows over the observed types of the following POS S or lexical item L:

$$H_{pos} = -\sum_S p_S log(p_S) \qquad H_{lex} = -\sum_L p_L log(p_L)$$

During the experiment, a POS-tagger analysed the strings in real time and triggered an intervention based on these entropy values, and a minimum requirement of 9 words (based on the mean length of all contributions). This manipulation produced a range of interventions with high, medium and low POS entropy, and, independently, high, medium and low lexical entropy.

Subjects and materials

The experiment was carried out on 16 pairs of students from Queen Mary University of London who were each paid £7.00 or given course credit for providing an hour of their time. The task was the *balloon task* – an ethical dilemma requiring agreement on which of three passengers should be thrown out of a hot air balloon that will crash, killing all the passengers, if one is not sacrificed. The choice is between a scientist, who believes he is on the brink of discovering a cure for cancer, a woman who is 7 months pregnant, and her husband, the pilot. This task was chosen on the basis that it is known to stimulate discussion, leading to dialogues of a sufficient length to enable an adequate number of interventions.

Subjects were seated at desktop computers in separate rooms, asked to input their e-mail address and username and given the task description. They were told that the experiment was investigating the differences in communication when conducted using a text-only interface as opposed to face-to-face, that the experiment would last approximately 45 minutes, and that all turns would be recorded anonymously for later analysis.

Analysis

Each intervention was annotated according to a number of factors. Firstly, whether or not there was a response

[1] This corpus is small, but extremely domain specific.

to the intervention during the timeout period. If there had been, the type of response was coded according to whether it was a compound contribution (CC), a clarification request (CR) or a yes/no response.[2] These are not mutually exclusive – example (1) is a CR constructed as a CC, and example (2) is a CC and a yes/no answer. The minimum POS entropy was 1.44, maximum 4.16, mean 3.27 (standard deviation 0.87); for lexical entropy those values are 5.59–8.14, mean 7.03 (s.d. 0.60).

(1) **B:** also surely the guy who knows how to ...
N: **fly?**
B: fly the baloon should know how to inscrease its height? *[DiET CCInd9 1277-80]*

(2) **J:** do you assess their value to society ...
Q: **in milliseconds yes =**
J: firstim with nick qne wuwi and susie - tom can explain how toise use the hot air balloon before he jumps *[DiET CCInd13 2048-51]*

The intervened turn was also annotated for whether it was potentially end-complete and could therefore be responded to as if it were a complete contribution. Antecedent end-completeness can be used as a proxy measure for pragmatic completeness, with 40 of the 241 truncated contributions appearing to end in a complete way.

The other major factor predicted to increase production of CCs was whether the subject under discussion was known to be shared. Lexical entropy gives us a measure of the predictability of the local context, with entities and concepts more or less predictable in certain sentential contexts because of the limited domain. However, it does not capture the potential effect on the predictability of local upcoming material of the *shared* context established in the course of any particular conversation between a specific pair of individuals. Each intervened contribution was therefore classified as either contributing to an ongoing topic of discussion, or introducing a new topic, as a loose measure of common ground.

Results

Of the 241 interventions, 171 elicited a response (71%). A GEE analysis with whether or not there was a response to the intervention as dependent variable[3] with POS and lexical entropy values as covariates, antecedent end-completeness as a fixed factor and participant as subject effect (goodness of fit QIC = 294.562; see table 1) showed a main effect of antecedent end-completeness such that responses were more likely in cases that could be considered complete on their own, showing that people are sensitive to TRPs.

[2]These response types were chosen on the basis of an examination of the response data.

[3]All models in this paper use a binary model with a logit link function and an independent correlation structure unless otherwise stated.

There was also an interaction effect of POS entropy by lexical entropy ($B = 0.237$, Wald-$\chi^2 = 5.893$, $p = 0.015$). This effect is illustrated in figure 2. Simple slopes analysis (Aiken et al., 1991) showed that responses are more likely in cases where both POS and lexical entropy were high (the highly unpredictable cases) than in cases where one or both levels of entropy were low.

	Model effects	
IV	Waldχ^2	p
Antecedent end-complete (Ant)	4.286	0.038*
Lexical entropy (Lex)	0.148	0.700
POS entropy (POS)	0.593	0.441
Ant × Lex	3.251	0.071
Ant × POS	2.546	0.111
Lex × POS	6.460	0.011*
Lex × POS × Ant	0.287	0.592

Table 1: GEE of response or not by lexical entropy, POS entropy and antecedent end-completeness

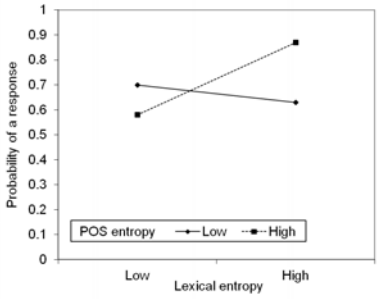

Figure 2: Marginal means of probability of a response by POS entropy × lexical entropy

Type of response

The results outlined above may conflate different effects which are specifically associated with different kinds of response. Analyses were therefore carried out separately on the different types of responses.

		Antecedent end-complete				
Response type		N	*%*	Y	*%*	Total
Yes/No	Y	20	*15*	12	*36*	32 *19*
	N	118	*85*	21	*64*	139 *81*
CR	Y	39	*28*	2	*6*	41 *24*
	N	99	*72*	31	*94*	130 *76*
CC	Y	62	*45*	10	*30*	72 *42*
	N	76	*55*	23	*70*	99 *58*
Total		138	*69*	33	*83*	171 *71*

Table 2: Response type

The breakdown of the 171 responses is shown in table 2.[4]

[4]Note that there were no differences in types of response according to which filler type was used ('...' or '...?').

Participants were more likely to produce a Yes/No response if the antecedent is end-complete ($\chi^2_{(1)} = 8.374, p = 0.004$), and they are also less likely to respond with a clarification request ($\chi^2_{(1)} = 7.201, p = 0.007$). There is no difference in the proportion of responses constructed as CCs based on whether the antecedent was end-complete or not, which is unexpected given the preference for expansions over completions in corpus studies (Purver et al., 2009).

CR responses

With the data filtered to responses only, GEE analyses on whether or not the response was formulated as a CR, with the POS and lexical entropy values as covariates and participant as subject effect (goodness of fit = 186.828) showed a main effect of POS entropy (see table 3). Greater syntactic predictability (lower POS entropy) increased the probability of the response being a clarification request.

	Model effects	
IV	Waldχ^2	p
Lexical entropy	2.207	0.137
POS entropy	5.135	0.023*
Lex × POS entropy	0.176	0.674

Table 3: GEEs CRs by lexical entropy, POS entropy and antecedent end-completeness

CRs are often formulated as CCs, as in (3) which is particularly true where the syntactic category of the next word was highly predictable (independently of lexical entropy). Of the 72 CCs, 21 occurred in syntactically predictable (low POS entropy) conditions with 12 of these also being CRs. Of the other 51 CCs, only 13 were also CRs (57% vs. 25%; $\chi^2_{(1)} = 6.575, p = 0.010$).

(3) **N:** i think susie because she is t ...
B: **a woman?**
N: ehe least important out of the three if you think about it ... dr nick is a doctor and could be really useful in the world *[DiET CCInd9 1214-7]*

CC responses

	Model effects	
IV	Waldχ^2	p
Antecedent end-complete (Ant)	1.951	0.162
Lexical entropy (Lex)	3.586	0.058
POS entropy (POS)	0.235	0.627
Ant × Lex	15.835	<0.001**
Ant × POS	0.018	0.894
Lex × POS	0.344	0.558
Ant × Lex × POS	0.005	0.945

Table 4: GEE of CCs by lexical entropy, POS entropy and antecedent end-completeness

GEE analyses on whether or not the response was formulated as a CC, with the POS and lexical entropy values as covariates, participant as subject effect and antecedent end-completeness as a fixed effect (goodness of fit = 234.351) showed an interaction between antecedent end-completeness × lexical entropy (table 4).

Simple slopes analysis shows that if the next lexical item is unpredictable then you are more likely to formulate your response as a CC if the antecedent is not end-complete. When the antecedent is end-complete (the solid line in figure 3), responses are more likely to be continuations in more highly predictable contexts (as in e.g. (4)), but when it is not end-complete CCs are more likely in the lexically unpredictable cases (as in e.g. (5)).

(4) **W:** I feel like we should be talking ...?
J: **about the prompt?**
W: about something important.
[DiET CCInd16 2846-9]

(5) **W:** nope we are not god we are ...?
M: **[M] and [W] ini lol we are [M] and [W] u fool lol so s just shut up npw please ad thank u for ur c kindeness**
W: not making dis di decision i knw we got bre spellintg werrorz man i r we even aloowed to talk type in slang? *[DiET CCInd6 929-32]*

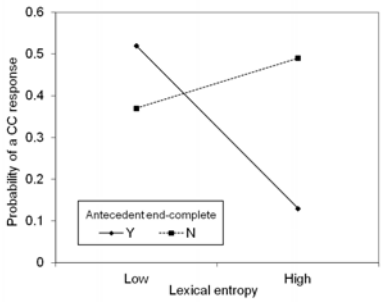

Figure 3: Marginal means of probability of a CC response by lexical entropy × antecedent end-completeness

Context

To test the hypothesis that CCs are more common where participants share information or common ground about the subject under discussion, planned post hoc analyses were carried out using the topic under discussion. Of the 241 intervened contributions, 170 were about an existing topic under discussion, whilst 71 introduced some new topic.

Participants were no more likely to respond if the turn was about the current topic or not; nor were they more likely to respond with a yes/no answer, or a clarification request. However, they were more likely to construct their response as a CC if it was about the current topic than if it was about something else (topic 59/121, 49% vs. Off-topic 13/50, 26%; $\chi^2_{(1)} = 7.519, p = 0.006$).

Adding topic to the GEE model with CC response as dependent variable (QIC = 227.895, table 5)[5] resulted in a three-way interaction effect of lexical entropy × POS entropy × topic.

	Model effects	
IV	Waldχ^2	p
Antecedent end-complete (Ant)	0.046	0.830
Topic	0.276	0.600
Lexical entropy (Lex)	2.545	0.111
POS entropy (POS)	0.018	0.892
Line number	2.361	0.124
Ant × Topic	0.381	0.537
Ant × Lex	3.435	0.064
Ant × POS	0.183	0.669
Topic × Lex	2.103	0.147
Topic × POS	0.281	0.596
Lex × POS	0.034	0.853
Ant × Topic × Lex	0.091	0.763
Ant × Topic × POS	0.005	0.946
Ant × Lex × POS	0.133	0.716
Topic × Lex × POS	8.635	0.003**

Table 5: GEE of type of CC responses by lexical entropy, POS entropy, antecedent end-completeness and topic

Exploring the interaction effect (figure 4) shows that in lexically unpredictable cases, which were syntactically predictable, participants were more likely to construct their response as a CC if they were talking about some topic which they had already been discussing, and which was therefore contextually salient.

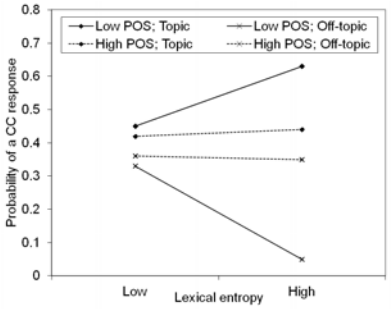

Figure 4: Marginal means of probability of a CC response by lexical entropy × POS entropy × topic

Discussion

These results offer some insights regarding the conditions influencing whether and how conversational partners respond to an incomplete utterance, and when they can and do construct those responses as continuations.

There is a response to 71% of the interventions, with this proportion affected by the predictability of the upcoming material. Perhaps counterintuitively, people are more likely to respond to unfinished contributions[6] if both syntactic and lexical items were unpredictable. This is not what we would expect if a simple model of levels of predictability were correct, as intuitively the most predictable cases ought to elicit the most responses. However, it is what we would expect if one of the drivers of human communication is in locally managing and resolving potential sources of misunderstanding (as in the interactive misalignment of Healey, 2008).

The main effect of potential completeness also demonstrates that people are more comfortable responding at all if the other person has reached a potential TRP – backing up findings from corpus studies (Purver et al., 2009) and conversation analysts assertion that people are sensitive to possible endings (Schegloff, 1996).

Compound contributions

Contrary to Hypotheses 1 and 2, continuations are not more likely at TRPs or syntactically predictable points. What is critical seems to be the actual and presumed accessibility of common ground. If the local content of what comes next is salient from the (presumed shared) context then people will produce completions. They do this by taking advantage of the syntactic structure of the antecedent, but syntactic predictability alone is not sufficient to prompt a completion.

A continuation response is more likely if the antecedent is complete but the next word is predictable (as in e.g. (4)) or if the antecedent is incomplete, suggesting that people complete where they can.

For the cases in which the antecedent is not end-complete, responses were more likely to be constructed as CCs in lexically unpredictable cases.However, if the next lexical item is highly predictable, then it can be interpreted as if it had actually been produced, as in (6). This result is not as surprising as it first appears as in a BNC corpus study (Howes et al., 2011), only 64% of end-incomplete contributions get continued, meaning that 36% never do. These are cases in which the local context is so predictable that it can be taken to be shared without the words themselves being produced.

(6) **T:** its not that fair on the girl doing th ...
H: exactly, you need to think of others and not be so selfish :P
T: study we should do lots of chatting although i doubt she'll read past the exercise what with it not being standardised etc *[DiET CCInd4 685-8]*

Context

The three-way interaction of POS entropy by lexical entropy by topic adds weight to the notion that what is critical to the production of a continuation in response

[5]The model also included line number as an additional covariate as it was found that participants were more likely to introduce a new topic later on in the conversation.

[6]This could be a genuine difference in text chat because of the availability of other cues in spoken dialogue, but we leave a discussion of this to one side.

to an incomplete utterance is the actual and presumed accessibility of common ground.

If the lexical item is unpredictable then syntactic predictability aids production of CCs in cases where the topic of the truncated contribution is shared, thus acting as a resource which helps frame the offered continuation as such. Syntax does not however help at all in cases where the topic is new so the gist of the contribution cannot be predicted and the predictability of the next word also offers no clues as to a plausible continuation.

This pattern of predictability corresponds to cases in which the high lexical entropy equates to lots of different words of a single type, as in the determiner case, rather than the high lexical entropy being associated with lots of different words of many different types (as with e.g. adverbs). This means that the syntactic category is highly constrained and the additional information associated with contextual salience can significantly narrow down an appropriate continuation.

Summary

This experiment, to the best of our knowledge the first to ever systematically attempt to induce continuations in an ongoing dialogue, shows that different types of predictability have different effects on what type of response participants produce to incomplete contributions, if any.

It shows that although syntax can be mobilised in constructing a response, it is not the crucial determinant of whether people construct their responses as continuations to the immediately preceding contribution. Participants make use of syntactic predictability only if the context is sufficiently constrained. Though people respect the constraints of the syntax, different points in the sentence do not cause greater difficulty in producing something that syntactically builds off a prior turn. However, that the grammar is a mutually available resource does not mean that it is used in the same way by all interlocutors, as evidenced by the finding that clarification requests are more likely, and more likely to be formulated as continuations, when the syntactic category of the upcoming material is more predictable, as these are cases where the syntax may be exploited to localise the source of a potential misunderstanding.

Another of the main findings is that people are sensitive to potential turn endings. These may be syntactic (in the antecedent end-complete cases) but they are not necessarily so. Some cases which appear to be syntactically incomplete can be responded to as if they are complete, provided that the continuation is highly predictable. If there are indeed cases which are interpreted as complete when they are not – as if the hearer is supplying the missing material internally, but does not necessarily produce it, this has implications for any grammatical or dialogue model. Incomplete syntactic strings must be not only successfully analysed, but also assigned potentially complete semantic representations.

The evidence from this experiment shows that when people are likely to produce CCs (or produce more CCs) is principally driven by common ground. They are possible (or more likely) when it is shared. How this is cashed out remains to be seen, however, it is apparent that some formal notion of context is crucial for a thorough understanding of CCs, especially if we are to ever hope to model them appropriately in a dialogue system.

References

Aiken, L. S., West, S. G., and Reno, R. R. *Multiple Regression: Testing and Interpreting Interactions*. Sage Publications, 1991.

Fernández, R. and Ginzburg, J. Non-sentential utterances: A corpus-based study. *Traitement Automatique des Langues*, 43(2), 2002.

Healey, P. G. T. Interactive misalignment: The role of repair in the development of group sub-languages. In Cooper, R. and Kempson, R., editors, *Language in Flux*. College Publications, 2008.

Healey, P. G. T., Purver, M., King, J., Ginzburg, J., and Mills, G. Experimenting with clarification in dialogue. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*. Boston, Massachusetts, 2003.

Howes, C., Purver, M., Healey, P. G. T., Mills, G. J., and Gregoromichelaki, E. On incrementality in dialogue: Evidence from compound contributions. *Dialogue and Discourse*, 2(1):279–311, 2011.

Lerner, G. H. On the syntax of sentences-in-progress. *Language in Society*, pages 441–458, 1991.

Lerner, G. H. On the "semi-permeable" character of grammatical units in conversation: Conditional entry into the turn space of another speaker. In Ochs, E., Schegloff, E. A., and Thompson, S. A., editors, *Interaction and Grammar*, pages 238–276. Cambridge University Press, 1996.

Purver, M., Howes, C., Gregoromichelaki, E., and Healey, P. G. T. Split utterances in dialogue: A corpus study. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009 Conference)*, pages 262–271. Association for Computational Linguistics, London, UK, 2009.

Schegloff, E. A. Turn organization: One intersection of grammar and interaction. In Ochs, E., Schegloff, E. A., and Thompson, S. A., editors, *Interaction and Grammar*, pages 52–133. Cambridge University Press, 1996.

Toutanova, K., Klein, D., Manning, C., and Singer, Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, pages 252–259. 2003.