# Simulating Cub Reporter Dialogues: The collection of naturalistic human-human dialogues for information access to text archives

**Emma Barker**[*], **Ryuichiro Higashinaka**[*†], **François Mairesse**[*], **Robert Gaizauskas**[*], **Marilyn Walker**[*], **Jonathan Foster**[**]

[*]Department of Computer Science/[**]Department of Journalism, University of Sheffield, UK
[†]NTT Communication Science Laboratories, NTT Corporation, Japan

**Abstract**

This paper describes a dialogue data collection experiment and resulting corpus for dialogues between a senior mobile journalist and a junior cub reporter back at the office. The purpose of the dialogue is for the mobile journalist to collect background information in preparation for an interview or on-the-site coverage of a breaking story. The cub reporter has access to text archives that contain such background information. A unique aspect of these dialogues is that they capture information-seeking behavior for an open-ended task against a large unstructured data source. Initial analyses of the corpus shows that the experimental design leads to real-time, mixed-initiative, highly interactive dialogues with many interesting properties.

## 1. Background and Motivation

For decades now it has been widely assumed that people seeking information from large-scale, electronic information resources want the option to communicate with the information access system as if they were communicating with another human, i.e. through a sequence of questions and responses forming a dialogue in which either party may take the initiative. Such information-seeking scenarios may be characterised by the nature of the information-seeking task – more or less focussed – and by the nature of the resource in which the information is being sought – more or less structured. Much work has been done on dialogues for information-seeking for focussed tasks against structured databases (Walker et al., 2002). Recently work has also been done on dialogues for information-seeking for less focussed tasks where the information source is a human expert (Stede and Schlangen, 2004; Denecke and Yasuda, 2005). Recent work has also begun to consider question asking and answering as potential means of access to unstructured, open domain text collections, but for single questions or short sequences of closely related questions only (Voorhees, 2005). This work has not yet advanced to the point of considering unconstrained dialogues for such an information-seeking scenario. However the ability to communicate via a dialogue with an information agent that has full knowledge of the content of a large text collection is clearly highly desirable (Burger et al., 2002).

The problem is that we do not know how to build systems that can do this, nor do we understand the character of naturalistic dialogues for access to open domain text collections, especially for less focussed and complex information-seeking tasks. As other work in human language technology has shown, a key step in advancing our knowledge is the creation of suitable language resources. Hence an important step towards understanding how to build dialogue systems for information access to large unstructured text collections is to gather a corpus of dialogues based on a realistic information-seeking scenario for a less focussed task. In this paper we address this issue, describing one such information-seeking scenario, the methodology we have adopted for creating such a corpus and the corpus that is currently under construction.

## 2. The Information Seeking Scenario: Background for Breaking News

Background information gathering in response to a new event – a scenario exemplified by the activities of journalists assembling background for breaking news stories – is one real world scenario where we find people seeking information from open domain text collections. Although in current practice background information seekers typically rely on standard information retrieval (IR) techniques to access information in text, we believe that this is a scenario which is of significant interest to dialogue studies. We explain this position by first describing the task in the news domain and the goals of people carrying out the task in more detail and then by summarising the key factors that, in our opinion, suggest this is a fruitful task for generating information-seeking dialogues of a type which have not yet been studied in detail, are richly interesting as dialogues, and are of significant practical importance.

There are a number of task contexts in which a journalist might require background. Some of these leave little obvious trace, such as when a journalist is preparing for an interview or inserting small amounts of background into a current story (e.g. *John Doe, president of FooBar Inc. and former world tiddlywinks champion, ...*). Others leave clearer traces, such as when, in support of a big story, a journalist is instructed to prepare a background fact-sheet to assist other journalists putting the current story in context (for example, a list of previous train crashes), or even to write a dedicated, so-called "backgrounder", which is an extended prose piece whose function is to contextualise the current event. In work reported elsewhere, which describes research into the application of Question Answering (QA), summarisation and Information Extraction (IE) technologies to support background gathering from news archives (Saggion et al., 2005), we focussed on the background writing task and carried out interviews with journalists, controlled observation of current practice and text analysis of around 50 archived backgrounders. This work has provided a better understanding of the goals of jour-

nalists seeking background information from a text archive and some insights into the nature of the process.

We observed two high level goals, each providing context for the current event. One is to provide simple descriptive information for entities or events that figure in the current story. For example, in background to a hurricane we see the proposition: *A storm can only be classified as a hurricane if its wind speed is faster than 73 mph*. A second is to find information about past events that can be used to frame the current event in a narrative that is both compelling and significant to the intended audience. Journalists describe this second goal as 'angle seeking' and while they may begin the task not knowing what the 'angle' may be, they have an expert understanding of the kind of information that needs to be examined in order to develop and support an angle. We can identify a number of types of information which are commonly used to provide angles in a news story, including: (1) chronological sequences of events; (2) possible explanations for a particular outcome; (3) interesting associations between groups of events or entities; (4) extreme or distinguished examples of similar events or entities; (5) information that places a current event or entity on a scale of similar events or entities.

We believe the following characteristics show why this scenario is a fruitful one to use in constructing a corpus of information-seeking dialogues and what makes it unique.

**Discovering new information from text archives: connecting entities and events** Given the goal of providing descriptive information for entities or events we can expect that background information-seeking dialogues are likely to include examples of simple factoid question-answering as well as exploration of unfamilar topics, as described in work on information-seeking chat (Stede and Schlangen, 2004), where the answer information is assumed to be explicit in the knowledge base (text archive, domain model) used to drive the dialogue.

We can contrast these examples of information-seeking with one where the information-seeker is looking to discover novel information by making connections between existing data. Journalists seeking background information are looking for information that could provide an angle for their story. While news archives often contain examples of background information to past news events, which might provide ready-made associations between events and entities, more commonly the journalist must identify the patterns in the archive himself, i.e. he has to actively make new associations between events or entities. We are not sure what kind of dialogues this will produce.

**Imperfect user model of information source** Background information-seeking is carried out against large text archives. In general a user will have only a poor model of the sort and extent of the knowledge held in such a resource. As a consequence part of the user's activity must be learning about the contents and limits of the resource he is searching against, not just about the topic of interest to him. This contrasts with information-seeking against structured databases where the user often has an accurate model of the type of information held in the database and the database may be complete with respect to the user's tasks.

**Inherently iterative** Because of the previous two characteristics – seeking to discover novel connections and imperfect user model of the information source – background information-seeking leads naturally to iterative information requests as the information-seeker learns about the topic, discovers connections, and learns about what archive contains. Dialogue is a natural mode of interaction in this situation.

**Complex/non-unique answers** Unlike factoid QA or information-seeking against structured databases, background information-seeking is unlikely to result in single, simple answers. Rather results are likely to be rich collections of related facts and different information-seekers carrying out the same task are likely to end up with different, but equally valuable results, because there are multiple possible angles.

**Unpredictable nature of dialogues** Since background information-seeking is a complex task without a precisely defined result, dialogues realising the process are likely to be more unpredictable and variable than those realising previously studied information-seeking tasks such as factoid QA against text collections or question answering against structured databases.

**Limitations of search engine technology** Background information-seeking is currently supported by conventional search engine technology. All of the journalists interviewed in this work have spontaneously voiced discontent with the limits of search engines for background information-seeking. After participating in the dialogues described below they were very enthusiastic about the utility of a dialogue-based interface to text archives that could support background information-seeking if such a system could be developed.

## 3. The Cubreporter Dialogue Scenario

Ideally we wish to capture information-seeking dialogues between two humans, one of whom has full knowledge of a news archive. However, no such individual exists. The closest we get in current practice are telephone conversations between a mobile reporter and a colleague in the news room, a "cub reporter", who acts as his proxy in gathering background, using a search engine over a news archive or a cuttings library. While this scenario suggests there is a real requirement for reporters to have remote access to background information via a spoken dialogue, the resulting dialogues are several steps removed from the sort of dialogues one would expect with an 'ideal' information provider. For example, the time required to get complete information is not feasible given news deadlines and thus the information provider typically relies on a handful of texts and is consequently not in a position to answer follow-on questions. Furthermore, such dialogues may contain long interruptions while the provider searches for relevant information or periods where the provider is simply reading aloud chunks from the archive.

To more fully approximate the situation in which dialogues are generated between an information provider (IP) who has fully assimilated the content of an archive and an information gatherer (IG) who is seeking background in the

In the years since the Second World War, Scotland has transformed itself from a nation of coal miners and shipbuilders into a modern, technology-based society. The cornerstone of this society is the Scottish higher education system. Scottish universities produce more graduates in the natural sciences, mathematics and computing on a per capita basis than Japan, the US or any country in Western Europe. Over recent years Scottish firms have been shown to spend far less on research than the UK average. In 1999, the Scottish executive created the Proof of Concept fund as a six-year, pounds 33m initiative. A funding gap was identified between the research activities in the university laboratory and those of a proven concept in which a commercial investor becomes interested. The fund, now in its fourth year, currently supports 120 groundbreaking projects worth over 19m. Companies that have started from projects receiving Proof of Concept funding include creative media company Virtual Clones Limited, which began life as a Glasgow University research project. The long-term advantages are two-fold: first, the research will be commercialised to the benefit of Scottish and international businesses. Second, participation in the project will produce a raft of highly skilled, commercially aware Scottish graduates to address the skills gap identified in this sector. Through these and other initiatives, Scotland is beginning to position itself within the international knowledge economy. Scottish Enterprise, Scotlands main economic development agency are providing the latest round of funding for the project. They say their key priorities of are to provide a range of high-quality services to:

- help new businesses get underway;
- support and develop existing businesses;
- help people gain the knowledge and skills they will need for tomorrow's jobs; and
- help Scottish businesses develop a strong presence in the global economy, building on Scotland's reputation as a great place to live, work and do business.

There were 124 applications for support but only 37 projects are sharing the 5.6m donated in the latest round of Proof of Concept awards. Among them are a Glasgow University project that screens for lung cancer by monitoring breath, and new third-generation mobile phone technology from Robert Gordon University in Aberdeen. Waging war on the common midge and reducing the amount of harmful gases belched out by grass-munching farm animals are two of the other 35 academic projects that are sharing the money. Meanwhile, Dundee University was granted three awards, worth 550,000 to support its concept for software mimicking human movements; to develop new drugs to cure or prevent potentially fatal fungal infections in cancer patients; and to develop a complete prototype system for a hearing device implanted in the ear. Abertay University received funding to build and test a software tool, designed to improve the competitiveness of Scottish digital games companies. It will also further develop its technology to speed up the purification of waste water from textile processing, paper making, food and drink processing and chemical factories. Edinburgh University got five awards, worth 950,000, one of which will see it develop a new technology to identify the presence of oil or gas underground, which it hoped would sustain North Sea oil production. To date the programme has made 120 awards and created 290 jobs. However, The Proof of Concept programme has come under intense criticism recently, following the collapse of Essient, the flagship first spin-out from Proof of Concept, which called in the liquidators last month after failing to raise second-round funding.

Figure 1: Sample background written for the breaking news story in Figure 2

news domain, we developed the following scenario.

A participant becomes expert in a small part of an archive (which we assume to be complete with respect to an information topic under consideration) via a controlled writing task, i.e. by writing a full backgrounder. We chose this task as a means to prime the participant for the role of IP, since writing provides a natural method for organising information and ideas. Having completed this task, the participant then plays the role of IP in a dialogue with an IG, a person playing the role of a reporter seeking information on the topic to inform his coverage of a new event.

We studied this scenario in a pilot experiment, which suggests the method facilitates the production of rich, naturalistic dialogues. We are now building a full corpus consisting of the archive texts used to inform the IPs, the corresponding background 'summary' texts produced by the IPs in the controlled writing task, and the dialogues between IPs and IGs. This corpus will allow us to explore research questions such as: (1) how different individuals seek or present information on different topics; (2) how one individual seeks or presents information on different topics; (3) how information presentation varies in style between speech and text; (4) system requirements for text processing on the source or background texts to support the behaviours of the information provider; (5) system requirements for a dialogue system targeting the source or background texts as a data source.

## 4. Experimental Method

Below we explain in more detail our method for simulating dialogues for information access to news archives.

### 4.1. Roles and Participants

We selected participants for the dialogues on the basis of skills and experience. For the role of IG, we recruited professional journalists from various media organisations including national news agencies, national broadsheets, regional newspapers, radio and magazines in addition to university lecturers in journalism who have prior experience working in one or more of the media listed above. Professionals have expertise in researching background for news, both in the context of writing tasks and in conversation with experts, colleagues and news archivists. This suggests they should comprehend the simulated "background for breaking news scenario" and the role of the IG with little difficulty. Moreover, we asked professionals to participate in the pilot study and they played the role of IG quite naturally and with purpose. They typically asked a range of questions on a topic, explored aspects of background through a series of connected questions, and queried new topics that arose.

For the role of information provider we sought participants with good verbal presentation skills and knowledge of researching and writing backgrounders since this experience should help them to (1) provide appropriate and useful responses in background for breaking news dialogues and (2) carry out a controlled background writing task using a collection of pre-selected archive news texts. Professional journalists, especially those working in broadcast news, typically have such skills. However, due to the practical constraint of obtaining large numbers of professionals, we asked journalism students with experience in writing news copy, background news gathering and radio journalism to participate in the role of IP.

### 4.2. Assimilating Information from Text

We prepare a participant for the role of IP in a cubreporter dialogue by asking them to carry out a controlled writing task using a small collection of archive texts. By using preselected sources – as opposed to allowing them to search for their own texts – we hoped to reduce the time spent by par-

ticipants in the search process and to exercise experimental control over the information source.

We designed the archive collection with two objectives: (1) to capture some of the characteristics of an open domain text collection, i.e. to include material from different news media, genres and topic areas; (2) to provide broad coverage of different types of background content for a given news story, in order to allow the IPs to view the current event from different perspectives. To meet these requirements we asked three independent researchers not participating as IPs or IGs to each carry out a lengthy archive search and to find up to 15 documents which they judged to be the best coverage of background information for a particular breaking news story (e.g. resignation of a politician, hurricane, etc.) and which would be suitable for use in a typical news wire backgrounder. To assist them in this task we provided a summary extract from a news wire story. Resources we used in the searches included the World Wide Web and specialised news archives, e.g. the Press Association archive, which contains news copy from agencies and national and regional newspapers. From this initial pool we removed obvious redundancy (i.e. identical sources) and then continued to deselect texts, aiming for a total of around 25 individual documents which in our judgement provided a good balance of background ingredients (e.g. material on different topics over time; definitional information; factoids; similar events; events leading up to the event). In the final resource we included nine sets of documents for nine news stories, three in each of three topic areas: natural disasters, company/person profiles; new investments.

We then asked IPs to write a 500 word news wire style backgrounder for one of the nine breaking news story using the corresponding prepared text collection. An example backgrounder written to support the breaking news story shown in Figure 2 is given in Figure 1. The IP task description included the same news wire summary extract that we used in the searches. Furthermore we provided brief directives to help them to focus their task, in the form of a "news editor's comments", e.g. for a story of a new commercial product release we wanted background to form a profile of the company. We encouraged participants to read each text briefly and to try and use multiple sources in the writing task. But otherwise they were free to use content as they wished. Finally, we wanted to obtain a record of the sources used in the task. Hence, upon completion, we instructed participants to indicate which sources they used for content in the course of writing their backgrounder.

### 4.3. Dialogue Roles

To each IP-IG pair we provided the details of a breaking news event and their dialogue roles in our scenario. The task of the IP was to imagine that he/she was employed on a news desk and, to the best of their ability, provide background information to callers, using the knowledge he/she had acquired during the writing task. This included both the information they had written about in the backgrounder and information they had not used but had remembered from their reading of the sources. To help prepare them for this task, shortly before a dialogue we asked IPs to spend 5-10 minutes reading over their background (written between 2 hours and 2 days in advance). We also gave them a copy of the background they had written for reference during the course of a dialogue if, for example, they wanted to check details for a fact or jog their memory. However we asked them to limit their use of the text, and try to avoid reading out loud from the written background as best they could.

TASK TWO

**New Investment: Proof of Concept Fund**
**Call Maggie on the news desk**
You are a newly appointed Scottish correspondent who is to cover the reaction to the news of a new round of investment in scientific research in Scotland. You plan to interview scientists, businessmen and politicians. Science funding is not a topic you are very familiar with and you need some **basic background**, especially information about the recent history of the funding body and investment in science in Scotland.

Breaking News Summary
Topic: 1 SCOTLAND Funding
Headline: MILLIONS INVESTED IN NEW RESEARCH PROJECTS
Byline: Scottish Press Association
Date: 2004-07-01

Scots scientists have secured millions of pounds of funding for a host of ground-breaking research projects, it was announced today. A total of 26 projects, from ways to improve cancer detection, tests for glaucoma and treatment to the development of a fish aerobics programme, have secured a share of a multi-million pound funding pot. The 4.5m Proof of Concept cash will benefit commercially-focused research projects taking place in universities, colleges and research institutes around the country. The latest round of funding, which is backed by Scottish Enterprise and the Scottish Executive, was announced today by deputy enterprise minister Lewis Macdonald... Todays announcement is the fifth round of support from the fund... Around 60 research posts will be created as a result...

Figure 2: Sample brief for mobile reporter, the Information Gatherer (IG)

For IGs, we prepared an "assignment brief" for the breaking news story they were to imagine they were covering. See Figure 2. This included details of their job in a news agency, e.g. "Scottish Correspondent"; details of an action they were planning to carry out, e.g. "interview scientists, businessmen and politicians". We also provided comments to indicate the focus for the story and the same breaking news story used by the IPs in the background writing task. Lastly, we asked IGs to initiate the dialogue and to continue to ask for information until he/she either reached the goal of recovering background for the assignment or believed he/she had exhausted the supply of relevant information from this source. We did not prompt either participant further since the pilot study suggested rich, mixed-initiative dialogues occur naturally.

### 4.4. Corpus Collection

The experimental design allows us to explore informally the effects of the variables: topic, seeker and provider. The final corpus will consist of 54 dialogues based on 9 IPs each writing 3 backgrounds, each of these in a different topic area (natural disasters, company/person profiles, new investments). For each brief two different IGs call each provider. An IG typically talks to three IPs, in each case following a background task assignment for a different topic.

## 5. Corpus Characteristics

There are six types of data in the corpus: task descriptions for IP and IG; archive source texts; background texts produced by IPs before a dialogue using a subset of the archive

| | Total | Mean | Min | Max | S.D. | Mean per topic | Mean per story |
|---|---|---|---|---|---|---|---|
| Number of of sources | 210 | | | | | 70 | 23.33 |
| Number of words | 131,278 | 625.13 | 70 | 7,262 | 616 | 43,759 | 14,586 |
| Number of sentences | 6,305 | 30.02 | 1 | 328 | 28.83 | 2,101 | 700.6 |
| Unique words | 56,877 | 270.84 | 53 | 1102 | 155 | 18,959 | 6,319 |

Table 1: Statistics of the archive source texts. Unless specified, counts are made over all source texts.

| | Total | Mean | Min | Max | S.D. |
|---|---|---|---|---|---|
| # of words | 8219 | 547.93 | 443 | 692 | 71.96 |
| # of sentences | 447 | 29.80 | 22 | 38 | 5.13 |
| Unique words | 3957 | 263.80 | 191 | 312 | 30.51 |
| Source texts used | 104 | 8.73 | 5 | 12 | 2.25 |

Table 2: Statistics of backgrounders.

| | Mean | Min | Max | S.D. |
|---|---|---|---|---|
| Dialogue duration (sec) | 412.23 | 207 | 721 | 128.98 |
| Average sources per IP | 8.5 | 5.0 | 10.5 | 1.73 |
| Overlap between IPs | 4.50 | 3 | 8 | 1.87 |

Table 3: Dialogues duration, variation between the average use of source texts by IP, and number of overlapping source texts used between pairs of IPs.

collection; records of the actual sources used in each writing task; spoken dialogue recordings; and full transcriptions of the dialogues.

### 5.1. Text Corpus

We employed GATE (Cunningham et al., 2002), a suite of linguistic processors including a tokeniser, a sentence splitter, a part-of-speech (POS) tagger, and a named entity recogniser, to perform a quantitative analysis of the source and background texts. Table 1 shows details of the source texts. There are 210 source texts, for a total of 131,278 words (6,305 sentences). There are 23.3 source texts on average for each topic area available to IPs to write a backgrounder and Table 2 shows that on average 8.7 texts are used per IP to write a story. We have 15 backgrounders and each of them has approximately 500 words. Table 3 shows that 4.5 of the source texts used are overlapping among the IPs, suggesting that some source texts are preferred over others in forming backgrounders.

### 5.2. Spoken Dialogues

We have collected 30 dialogues, whose duration is on average 6:52 minutes with a standard deviation of 2:09 minutes. We are in the process of transcribing and analyzing these dialogues.

A primary motivation with our experimental design was to collect a corpus of information-seeking conversations against a textual information source that would have the properties of real-time interactive dialogue. We particularly wished to avoid collecting dialogues with: (1) long pauses in which the IP actively searched for background information in the text archive; and (2) long dialogue segments in which the IP read aloud from textual sources. Our pilot experiment suggested that our experimental design should achieve this. Observations while the corpus was being collected, and initial analyses of transcripts show that the dialogues have the desired properties (Schegloff, 1982; Clark and Schaefer, 1987). For example:

- There are interruptions, overlapping speech, latched speech, other completions, sentence fragments, and few long pauses;
- Referring expressions and other aspects of meaning are often collaboratively constructed;
- There is frequent use of pronominals, textual deictics and other indexical expressions;
- The dialogues are mixed initiative. Although the IG and the IP have particular roles, initiative is passed between them with the IG taking control and interrupting to pursue particular angles or aspects of the information provided, and the IP often volunteering unasked for information, or answering different but related questions.

We can illustrate some of these characteristics with examples from a dialogue for the task in Figure 2, with the full background text prepared by the IP in Figure 1.

The dialogue excerpt in Figure 3 illustrates iterative questioning and responding, i.e. where the IG interrupts (IG1) to ask a question about content in IP1. The dialogue segment from IG1 to IG5 shows the collaborative construction of a referring expression *June 2004*, where the IP and IG work together to ascertain the date when a particular event took place. It also contains a large number of deictics.

| | |
|---|---|
| IP1: | ...but the liquidators were called in fairly recently |
| IG1: | When was this, liquidators? |
| IP2: | Sorry |
| IG2: | When was this? |
| IP3: | Well it says last month |
| IG3: | Which year, what year was it, this year? |
| IP4: | Last month, it's 2005 now according to this. |
| IG4: | So it's last year I can say? |
| IP5: | Yeah so it's like June 2004 |
| IG5: | Brilliant so June 2004. |

Figure 3: Sample dialogue 1.

Figure 4 illustrates angle seeking by the IG, who wants to know of any big achievement or project successes following a previous funding award, and later, if any of the projects have been concerned with 'animal science'. The IP responses to each of these requests illustrate cooperative dialogue. The IP states that she does not have available information (IP1), and she indicates the scope of what she does know. In IP2, the IP takes the initiative to offer an alternative type of success. Of particular note is the fact that the IP has generalised over the initial information request,

and offered a related proposition from a broader class of possible answers. A similar demonstration of reasoning via relaxation to a broader class is shown in IP4, where the IP acknowledges that a project on midges might not count as animal science, and a further exchange occurs to clarify in what class the project does fall.

| | |
|---|---|
| IG1: | Uh, okay, what's their biggest achievement, is there anything about how they've ordered money for anything kind of high profile or have you found anything, have you found any success attached to...? |
| IP1: | No not particularly, I sort of know more about how much money has been given to various people and what for. As opposed to successes. |
| IG2: | Okay. |
| IP2: | But I do know one good thing, is that it's created 290 jobs. |
| IG3: | 290 jobs in the space of six years, jobs okay. |
| IP3: | And 120 projects have had awards over that time. |
| IG4: | Oh brilliant, over the six years. Has there been anything to do at all with animal science? |
| IP4: | Animal science, well there's, I don't know if they count as animals actually but there is a project against midges, it actually says waging war on the common midge, and reducing the amount of harmful gas produced by munching farm animals |
| IG5: | But it's not sort of directly animal science |
| IP5: | No it's more sort of environmental I think |

Figure 4: Sample dialogue 2.

In Figure 5, the IP takes the initiative (IP3 to IP7), to provide further information on the role of the funding body, speculating on the meaning of the information in her response. There is also a reformulation in IG7, where the IG reframes the conceptualization of the funding body, and gets the IP's assent. We hope to quantify the extent to which these various phenomena occur in future work.

| | |
|---|---|
| IG1: | Hmmm and what does it, does it kind of dole out cash to scientific bodies so that they can research certain things? |
| IP1: | Humm basically yeah what they do err what they do is err bodies and universities apply err for money |
| IG2: | Yes |
| IP2: | For for awards from it |
| IG3: | Oh |
| IP3: | Hmmm and then they allocate it, I think they must be allocated depending on what they want to do with the money |
| IG4: | I see |
| IP4: | Hmmm because for this one there are 124 applications |
| IG5: | Ok |
| IP5: | But only 37 of the applicants actually got money |
| IG6: | Oh right |
| IP6: | So they must judge them on something but I don't know for sure what they judge them on |
| IG7: | So it's a bit, ok I get you, so it's like an award body |
| IP7: | Exactly |
| IG8: | For scientific communities |
| IP8: | Yeah |

Figure 5: Sample dialogue 3.

## 6. Conclusion

We have described an experiment to construct a corpus of naturalistic dialogues between an information-seeker with a partially specified information need and an information-provider, where the latter is conceived of as having full knowledge of a large unstructured open domain text collection. In future work we plan to use this corpus to explore: (1) how different individuals seek or present information on different topics; (2) how one individual seeks or presents information on different topics; (3) how information presentation varies in style between speech and text; (4) system requirements for text processing on the source or background texts to support the behaviours of the information provider; (5) system requirements for a dialogue system targeting the source or background texts as a data source.

## Acknowledgments

This work has been supported by the UK EPSRC Grant GR/R91465, and a research collaboration grant from NTT Communication Science Laboratories to the Cognitive Systems Group at University of Sheffield. We are also grateful to the journalists and journalism students who participated in our experiment.

## 7. References

J. Burger, C. Cardie, V. Chaudhri, R. Gaizauskas, S. Harabagiu, D. Israel, C. Jacquemin, C-Y. Lin, S. Maiorano, G. Miller, D. Moldovan, B. Ogden, J. Prager, E. Riloff, A. Singhal, R. Shrihari, T. Strzalkowski, E. Voorhees, and R. Weischedel. 2002. Issues, tasks and program structures to roadmap research in question & answering (q&a). Technical report. www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc.

H. Clark and E. Schaefer. 1987. Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2:19–41.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. 40th ACL*.

Matthias Denecke and Norihito Yasuda. 2005. Does this answer your question? towards dialogue management for restricted domain question answering systems. In *Proc. 6th SIGDial Workshop on Discourse and Dialogue*.

H. Saggion, E. Barker, R. Gaizauskas, and J. Foster. 2005. Integrating NLP tools to support information access to news archives. In *Proc. 5th International Conference on Recent Advances in Natural Language Processing RANLP-2005*.

E. A. Schegloff. 1982. Discourse as an interactional achievement: Some uses of "uh huh" and other things that come between sentences. In D. Tannen, editor, *Analyzing Discourse: Text and Talk*, pages 71–93. Georgetown University Press.

M. Stede and D. Schlangen. 2004. Information-seeking chat: Dialogues driven by topic structure. In *Proc. the 8th Workshop on Semantics and Pragmatics of Dialogue*.

E. Voorhees. 2005. Overview of the TREC 2004 question answering track. In *Proc. 13th Text Retrieval Conference (TREC 2004)*. NIST Special Publication 500-261.

M. A. Walker, A. I. Rudnicky, R. Prasad, J.Aberdeen, E. O. Bratt, J. S. Garofolo, H. Hastie, A. N. Le, B. Pellom, A. Potamianos, R. Passonneau, S. Roukos, G. A. Sanders, S. Seneff, and D. Stallard. 2002. DARPA communicator: Cross-System Results for the 2001 Evaluation. In *Proc. ICSLP*, pages 269–272.