THE HCRC MAP TASK CORPUS: NATURAL DIALOGUE FOR SPEECH RECOGNITION

Henry S. Thompson[1,2,3]
Anne Anderson[1,5]
Ellen Gurman Bard[1,3,4]
Gwyneth Doherty-Sneddon[1,5]
Alison Newlands[1,5]
Cathy Sotillo[1,4]

1: Human Communication Research Centre
2: Department of Artificial Intelligence
3: Centre for Cognitive Science
4: Department of Linguistics
University of Edinburgh
2 Buccleuch Place,
Edinburgh, EH12 5BB
SCOTLAND

5: Department of Psychology
University of Glasgow
56 Hillhead Street
Glasgow, G12
SCOTLAND

hthompson@edinburgh.ac.uk

ABSTRACT

The HCRC Map Task corpus has been collected and transcribed in Glasgow and Edinburgh, and recently published on CD-ROM. This effort was made possible by funding from the British Economic and Social Research Council.

The corpus is composed of 128 two-person conversations in both high-quality digital audio and orthographic transcriptions, amounting to 18 hours and 150,000 words respectively.

The experimental design is quite detailed and complex, allowing a number of different phonemic, syntactico-semantic and pragmatic contrasts to be explored in a controlled way.

The corpus is a uniquely valuable resource for speech recognition research in particular, as we move from developing systems intended for controlled use by familiar users to systems intended for less constrained circumstances and naive or occasional users. Examples supporting this claim are given, including preliminary evidence of the phonetic consequences of second mention and the impact of different styles of referent negotiation on communicative efficacy.

1. INTRODUCTION

The HCRC Map Task corpus has been collected and transcribed in Glasgow and Edinburgh, and recently published on CD-ROM (HCRC 1993). This effort was made possible by funding from the British Economic and Social Research Council.

The group which designed and collected the corpus covers a wide range of interests and the corpus reflects this, providing a resource for studies of natural dialogue from many different perspectives.

In this paper we will give a brief summary of the experimental design, and then concentrate on those aspects of the corpus which make it a uniquely valuable resource for speech recognition research in particular, as we move from developing systems intended for controlled use by familiar users to systems intended for less constrained circumstances and naive or occasional users. Some preliminary results of work on the phonetic consequences of second mention and on the impact of different styles of referent negotiation on communicative efficacy will also be presented.

2. CORPUS DESIGN AND CHARACTERISTICS

2.1. The Task

The conversations were elicited by an exercise in task-oriented cooperative problem solving. The two participants sat facing one another in a small recording studio, separated by a table on which sat back-to-back reading stands. On each stand was a schematic map, each visible only to one participant. Each map consisted of an outline and roughly a dozen labelled features (e.g. "white cottage", "Green Bay", "oak forest"). Most features are common to the two maps, but not all, and the participants were informed of this. One map had a route drawn in, the other did not. The task was for the participant without the route to draw one on the basis of discussion with the participant with the route.

2.2. Experimental Design

Using an elaboration of a design developed over a number of years (see e.g. Brown, Anderson et al. 1983), we recorded 128 two-person conversations (each talker in four conversations), employing 64 talkers (32 male, 32 female), almost all born and raised in the Glasgow area, speaking with an educated West of Scotland accent. High quality recordings were made using Shure SM10A close-talking microphones, one talker per channel on stereo DAT (Sony DTC1000ES).

The experimental design is quite detailed and complex, allowing a number of different phonemic, syntactico-semantic and pragmatic contrasts to be explored in a controlled way. In particular, maps and feature names were designed to allow for controlled exploration of phonological reductions of various kinds in a number of different referential contexts, and to provide a range of different stimuli to referent negotiation, based on matches and mis-matches between the two maps.

Among the independent variables in the design were:

- Eye-contact—in half the conversations, the participants could see one another's faces, in half, they could not.
- Familiarity—in half the conversations, the talkers were acquaintances, in half, strangers.
- Task role—Each talker participated in four conversations, two as Instruction Giver (the one with the route) and two as Instruction Follower (the one trying to draw it)

For a complete description of the experimental design, see Anderson, Bader et al. (1991).

2.3. Corpus Characteristics

Subjects accommodated easily to the task and experimental setting, and produced evidently unselfconscious and fluent speech. The syntax is largely clausal rather than sentential; showing good turn-taking, with relatively little overlap/interruption. The total corpus runs about 18 hours of speech, yielding 150,000 word tokens drawn from 2,000 word form types. Word lists containing all the feature names were also elicited from all speakers, along with a number of 'accent diagnosis' utterances.

The acoustic quality of the recordings is good but not outstanding—in particular, stereo separation is not perfect, in that it is often possible to detect the voice of one talker very faintly on the other talker's channel. A very modest amount of rumble and other non-specific background noise is occasionally detectable.

3. THE TRANSCRIPTIONS

The transcriptions are at the orthographic level, quite detailed, including filled pauses, false starts and repetitions, broken words, etc. Considerable care has been taken to ensure consistency of notation, which is thoroughly documented. Although the full complexity of overlapped regions has not been reflected in the transcriptions, such regions *are* clearly set off from the rest of the transcripts. Transcripts are connected to the acoustic sampled data by sample numbers marked every few turns.

Text Encoding Initiative-compliant SGML markup is used, both within transcripts to indicate turn boundaries and for other metatextual purposes, and also in separate corpus header and transcript header files, but this was done in a manner designed to make accessing the transcripts as plain text very easy.

We also used a very light-weight non-TEI markup for textual annotations, to mark such things as abandoned words, letter names, filled pauses and editorial uncertainties.

A brief extract from a transcript is given below as Figure 1, illustrating various aspects of the transcription, including the tags `u` for utterance, `sfo` for speech file offset, `bo` for begin overlap and `eo` for end overlap, as well as the `le` microtag for a letter name.

```
<u who=G n=3>
<sfo samp=107715>
<bo id=o75a>
About half an inch above it, we've
got an {le|x} marking start.  Have

<u who=F n=4>
<sfo samp=208987>
Yes.

<u who=G n=5>
you got that?
<eo id=o75a>
```

Figure 1. Extract from a Map Task Corpus transcript

4. THE CD-ROMS

The published version of the corpus occupies 8 CD-ROMs, and contains:

- a complete set of transcripts;
- 20KHz sampled versions of both channels of the associated speech for all the conversations;
- for each talker sampled audio for an accent diagnostic passage and a scripted reading of a list of all the feature names from the map;
- images of the maps employed;
- documentation;
- UNIX™ tools for linking the spoken and written material and other manipulations of the corpus materials.

Preparation of the corpus for publication was a much larger task than we had expected, and is described in some detail in (Thompson & Bader, 1993).

5. IMPLICATIONS FOR SPEECH RECOGNITION

5.1. High-quality unscripted dialogue

Recorded collections of natural conversation are not new—not only do many linguists have a drawer full of tapes of dinner table or staff room talk, but also more systematic and extensive collection efforts have been carried out on several occasions as part of major reference corpora building projects. But with no exceptions we are aware of, all such material is of highly varying acoustic quality, and is rarely if ever suitable for extensive computational processing.

On the other hand, to date the large development corpora collected and used to such good effect by the speech recognition community, although of a very high standard acoustically, have been exclusively monologue, and until very recently exclusively scripted.

Thus the Map Task corpus occupies a hitherto vacant position in corpus design space—it is natural, unscripted dialogue recorded to a standard suitable for digital processing. We hope the widespread availability of such a resource will help to stimulate a change in the way phonology, morphology, syntax and semantics are pursued parallel to the change which has already occurred in phonetics, that is, a change from theory development dependent on small amounts of data, often constructed by the theorist, to theory development dependent on, indeed immersed within, a large amount of naturally occurring data.

Note that this methodological change is, or at least ought to be, independent of meta-theoretical disposition, and in particular the above remarks are *not* meant to imply a bias in favour of stochastic or self-organising theoretical frameworks.

5.2. Syntax

There is modest controversy brewing about the relation between spoken and written language, particularly in a highly literate language/culture context such as obtains for English. It has been argued (see e.g. Miller 1993) that the grammar of spoken English is qualitatively different from that of written English, and demands separate treatment.

In so far as the progress of speech recognition from relatively constrained interaction situations and relatively constrained language will depend on grammars and/or models of natural English conversation, the resource provided by the Map Task has an obvious rule to play.

5.3. Prosody

It has long been assumed that there is a mutually informing relationship between prosody and discourse structure. The simple goal-oriented nature of the Map Task conversations, and the ease with which quite local, short-term goals can be identified in terms of the part of the route in question at any given time, means that the corpus provides an excellent base at attempting to explicate this relationship in some detail. Work has begun on relating the inventories of intonation on the one hand and moves within conversational games on the other, with initially encouraging results (Kowtko, Isard and Doherty 1992).

As in the case of syntax, we would hope that widespread provision of the corpus will enable comparative exploration of the numerous theories of discourse structure, prosody and their relations now being suggested.

5.4. Fast speech rules

The names associated with the landmarks drawn on the maps were designed *inter alia* to provide opportunities for various forms of phonological modification, in particular t-deletion ("vast meadow"), d-deletion ("reclaimed fields"), glottalisation ("white mountain") and nasal assimilation ("crane bay"). Furthermore, on each map one such name would be paired with another, similar name, with the intention of assessing the impact of the (putative) necessity of contrastive stress ("crane bay" vs. "green bay"). The availability in the corpus of citation form pronunciations be each speaker will provide a very useful baseline for studies in this area.

		Whole Corpus	Eye Contact	No Eye Contact
Word form types		2070	1553	1558
Word tokens	All turns	152298	69762	82536
	Instruction Giver	104828	48361	56467
	Instruction Follower	47470	21401	26069
	per conversation	1190	1090	1290
Turns	All turns	21251	9513	11738
	Instruction Giver	10678	4777	5901
	Instruction Follower	10573	4736	5837
	per conversation	166	149	183

Table 1. Summary corpus statistics

5.5. The role of eye contact

Not surprisingly, there are obvious gross effects on the conversations of the difference between the eye-contact and no-eye-contact conditions. The no-eye-contact conversations contained 22% more turns on average, but only 18% more words, i.e. more turns, but each fewer words per turn. This is presumably because of the increased need for frequent back-channel confirmations in the no-eye-contact condition.

The overall statistics for word tokens and turns are as given in Table 1. The implications of the language differences induced by the presence or absence of eye-contact are clearly significant for a range of different potential speech technology applications. See (Boyle, Anderson & Newlands, in press) for more details.

6. PRELIMINARY RESULTS

6.1. Second Mention

The duration and/or (excerpted) intelligibility of different tokens of a word uttered by the same speaker have been shown to depend on the availability of information outwith the word's acoustic shape which might help listeners to recognize it. In the context of extended discourse, this means word tokens are less intelligible when they refer to Given entities.

On the face of it, the tendency to produce degraded tokens where they are redundant seems wonderfully cooperative, in the Gricean sense of the term: when there is previous relevant material, intelligibility is reduced. The less intelligible repeated tokens are in fact helpful to listeners, for they make better prompts to earlier discourse material, either because they signal listeners to associate the word's meaning with some entity already established in a discourse model (Fowler and Housum, 1987) or because such stored information must be called into play for successful on-line word recognition (Bard et al., 1991).

The difficulty is that degraded tokens are not restricted to contexts in which the listener can recover the conditioning information. Using the Map Task corpus, we have begun to investigate how far speakers' adjustment of intelligibility is egocentrically rather than cooperatively based, that is, how far the speaker's own relevant knowledge provides his/her model for what the listener knows.

We have found the expected loss of intelligibility for excerpted second mentions as against both first mentions and citation forms. Interestingly, we found that the coreferential repetition effect found for monologue holds in dialogue: it doesn't matter who utters the word first. When it comes to the second mention of an entity either speaker may reduce intelligibility. This suggests that dialogue participants maintain a common record of textually evoked given entities.

It would also appear that once an entity is textually evoked there is no further effect of visual information. That is, it doesn't matter whether the listener or speaker can see the object they're referring to.

Also relevant is a significant intelligibility loss we found in mentions which are only 'second' for the speaker, because the relevant feature was first mentioned not in the current conversation, but in a previous conversation *with a different listener.*

Thus on the basis of our investigations to date it would appear, somewhat surprisingly, that speakers reduce articulatory effort on a purely egocentric basis, without regard to listeners' ability to share the contextual conditioning this implies.

6.2. Efficacy of New Item Introduction

The Map Task corpus presents an excellent opportunity for examining how speakers introduce new items into a discourse. Moreover, because we can measure to overall communicative effectiveness of a conversation by reference to the accuracy of the resulting map, we can go further and attempt to assess the value of particular item introduction strategies.

Definite versus indefinite article is almost certainly too simplistic a starting point for investigating this issue. This is born out by a tabulation of new item introduction over half the corpus, as shown in Table 2. 'Question' introductions are those in which the speaker queries the existence of the referent, e.g. "Right you got an extinct volcano?".

	Non-Question Introductions			Question Introductions		
Articles:	None	Definite	Indefinite	None	Definite	Indefinite
IG. Introductions	0.54	2.75	0.35	1.5	1.48	3.29
IF. Introductions	0.67	0.87	1.2	0.32	0.34	0.46

Table 2. Mean number of introductions per dialogue by form of introductions used by instruction givers (IG) and instruction followers (IF).

Overall definites and indefinites appeared with equal frequency.

If we look at listener's responses to the introduction of items they don't have on their map, we see a significant correlation of informative responses ("I haven't got an extinct volcano") with question introductions, but not with indefinite article usage as such. Also, using the accuracy of the route drawn as a measure of communicative efficacy, we found a significant correlation between use of question introductions by the IG and route accuracy. There is an independent correlation between informative IF responses and accuracy. See (Anderson & Boyle, in press) for more details.

7. CONCLUSION

The HCRC Map Task corpus has been designed to allow investigation of a range of issues relevant to both psychological models of human language production and comprehension and to speech technology, especially as the focus on effort switches to more natural, unconstrained speech. Preliminary results of studies in several areas provide encouraging evidence that the corpus will indeed yield valuable insights.

REFERENCES

1. Anderson, A. H., M. Bader, E. G. Bard, E. H. Boyle, G. M. Doherty, S. C. Garrod, S. D. Isard, J. C. Kowtko, J. M. McAllister, J. Miller, C. F. Sotillo, H. S. Thompson and R. Weinert. "The HCRC Map Task Corpus", *Language and Speech* 34(4), 1991, 351–366.
2. Anderson, A. H. and E. Boyle. "Forms of introduction in dialogue, their discourse contexts and communicative consequences", *Language and Cognitive Processes*, in press.
3. Bard, E.G., L. Cooper, J. Kowtko and C. Brew. "Psycholinguistic studies on the incremental recognition of speech: A revised and extended introduction to the messy and the sticky", University of Edinburgh: Centre for Cognitive Science DYANA Report R1.3.B, 1991.
4. Boyle, E. H., A. Anderson and A. Newlands. "The effects of visibility on dialogue & performance", *Language and Speech*, in press.
5. Brown, G., A. Anderson, G. Yule and R. Shillcock. *Teaching Talk,* Cambridge, U.K.: Cambridge University Press, 1983.
6. Fowler, C. and J. Housum. "Talkers' signalling of 'new' and 'old' words in speech and listeners' perception and use of the distinction", *Journal of Memory and Language*, 26, 1987, 489–505.
7. Human Communication Research Centre. *HCRC Map Task Corpus*, Edinburgh, U.K.: HCRC, 1993.
8. Kowtko, J., S. Isard and G. Doherty. *Conversational games within dialogue*, University of Edinburgh: HCRC Technical Report RP-31, 1992.
9. Miller, J. "Spoken and written language: language acquisition and literacy". in R. Scholes, ed., *Linguistics and Literacy,* Lawrence Erlbaum, 1993.
10. Thompson, H.S. and M. Bader. *Publishing a Spoken and Written Corpus on CD-ROM: The HCRC Map Task Experience.*, University of Edinburgh: HCRC Technical Report, 1993.