Proceedings of the Institute of Acoustics

FUNDAMENTAL RESEARCH UNDERLYING THE DESIGN OF AN AUTOMATED DICTATION SYSTEM

James Monaghan

University of Hertfordshire, Speech & Language Technology (*SLANT*), Hatfield, Herts, England, AL10 9AB

1. INTRODUCTION

SLANT has made its reputation since its inception at the Hatfield Polytechnic (now the University of Hertfordshire) both in the development of practical speech recognition-based systems and in basic research. Practical systems include the ISDIP Speech-Driven Word-Processor, which won a RITA award and was short-listed for a BCS prize, and the *SLANT* generalised speech interface, which is at present being designed and developed for use in a classroom environment by students with various motor handicaps[1]. This work is part of the Computer Aided Learning Enhancement Project (CALE), and is being partly supported by a substantial financial donation from Save and Prosper Educational Trust. The basic research work in the area of speech and language technology which supports this work has been concentrated on human factors issues associated with the use of the speech technology in real-world situations, as well as on the phonetic and linguistic underpinning of various types of text-production. It is on this latter area that I wish to concentrate in the present paper.

In the various linguistic technologies such as speech I/O and dialogue design for human-computer interaction it has become customary to distinguish between speech and natural language processing. This is obviously useful in areas where signal processing is particularly in the foreground. However, it is important to remember that speech is, from a linguistic point of view, just an outward instantiation of natural language, and a lot of the redundancy underlying the human understanding of spoken language depends on this. In other words, when you listen to what I am saying in my paper, your experience of the English language will give you lots of expectations about what I am going to say next. This built-in redundancy saves you having to attend to every single word and this is especially useful when the acoustics are less than perfect. I will, therefore, use the term spoken language instead of speech to emphasize this orientation. Most natural language is to be found in the spoken form and this is true even in an environment like the office where a large number of activities and processes involve written texts - a fact to which I will return below.

2. THE OFFICE AS AN INFORMATION ENVIRONMENT

Any instance of spoken language must take place in a context - a meaningful set of situational parameters which have to be taken into account in order to allow the interactants to create and exchange meaning. Conversely, the language used helps to create and define this context. It is important, however, not to confuse this abstract notion of the context with the mere physical surroundings. The same physical surroundings during the office party, or at 5 a.m. when the cleaners are in, is a place where completely different sets of linguistic norms apply in comparison with the

FUNDAMENTAL RESEARCH UNDERLYING THE DESIGN OF AN AUTOMATED DICTATION SYSTEM

working office where the office staff are fulfilling the tasks for which they are paid. The office, in our sense then, is an abstract organisational context designed to enable human interaction to take place whose overall purpose is to generate, process, store and disseminate information. The core of the interaction in the office as defined is unmediated person-to-person, as distinct from, say, a telesales environment or a software development lab. Tools appropriate to the underlying goals of the organisation are deployed to allow the most creative and profitable use of the mental resources of the human beings who work there, and these become part of the context. Inappropriate equipment tends to replace task-directed human activity with machine-driven behaviour, which may be tangential to the overall organisational goals. Struggling with inappropriate equipment is not fulfilling the tasks which both the workers and the equipment are there to solve. And the faster and more powerful the equipment is, the more easily the human users can be overwhelmed by it. Therefore, it is important to make an assessment of what users do without equipment and then to compare this with what can be done with what the technology allows.

To a depressingly large extent the dynamic behind many of the innovations in this area in the 80s was based on speeding up and/or combining aspects of existing equipment, rather than on enabling the underlying human resources to operate to their optimal capacity. Inevitably this leads to constraints on the pursuit of the overall organisational goals. As Diaper points out [2]:

> 'Current practice in a task is frequently tied to the existing technology employed in the task and it is therefore difficult to produce a creative, novel solution to system design based on such methods.

The answer to this problem is, of course, not simple, because whatever technology is provided will, in time, produce a new creativity which will, in turn, produce new requirements and assumptions about the potential for new design.

There are several useful techniques which support the design process without over-dependence on existing technology and practice. One, which also functions to reduce the need for early prototypes in the design life cycle is *scenario building*[3]. Scenarios cannot, however, be constructed in a vacuum, but must be built from a body of knowledge of real organisational goals, tasks and roles. These in turn need firm justification and cannot rely on impressionistic criteria. One very successful basis for these constructs in the case of many information processing tasks is to derive them from the language associated with them [4]. Such a procedure is, however, initially very time consuming because of the data collection, analysis and modelling involved. I propose to sketch the application of such a methodology for one important aspect of the office environment.

As studies have progressed it has become evident that the identification of user requirements is central to the design of future systems which must be both efficient in terms of task accomplishment and acceptable to the end user. These requirements can only be identified through the medium of natural language, and the collection and analysis of spoken language data has become an important part of the modelling process

which is the aim of task analysis in the office context. The spoken language data used for such studies has, however, been selected so as to fall primarily into three major categories - interview (structured or otherwise), concurrent verbal protocol and walkthrough protocol (task description by the task performer either during or after the task) [5], and has been analysed only to the extent necessary to support the structures already singled out for task analysis. Natural language has been used in this way simply to support the task analytical approach, rather than to provide an independent but related body of knowledge which can be used to test, enrich and elucidate the structures of task analysis. Even in studies where the primary focus has been on the role of language in office systems, the analyses have been at a very high level [6,7], and have largely neglected the body of knowledge which is relevant to the office context and which has been developed within the field of linguistics over the past twenty years. Similarly, the direct speech equivalents of certain language based tasks - e.g. taking dictation, setting up meetings, diary management - have been omitted from the analysis and modelling process.

Over the same period, a body of knowledge has been developed in linguistics about how language-based activity is structured and thus predictable. For example, work in grammar [8], intonation [9,10,11], turn-taking and topic management in telephone conversations [12,13], discourse structuring [14,15,16], and speech recognition [17] show how cues at many different levels are used by humans to predict what is going to be said next. These techniques provide a set of analytical tools for the identification of low-level features of language, which work together to signal different boundaries and structures in the language corresponding to similar segmentations in the ongoing activity. Applications immediately relevant to office automation are areas like the design of vocabularies in speech recognition systems, and the automation of tasks like dictation and diary management.

3. DICTATION

One area particularly appropriate for automation is dictation. This paper will report on the early stages of research into the design and development of an automated dictation system, currently being carried out within the *SLANT* team. The paper will describe and discuss the fundamental problems which must be addressed in such a project, and will suggest solutions to allow implementation of an automated system which will be able to cope with those problems.

The differences between what is actually said during a dictation session and what appears in the final document will be discussed and illustrated by text handouts, and audio taped examples from a real dictation session will be used to illustrate specific points. Particular attention will be paid to filled and unfilled pauses in the raw data, and evidence from the audio tapes will be given to support the hypothesis that their function is to signal the structure of the spoken material in terms of both the grammar of the sentence and the grammar of the discourse.

FUNDAMENTAL RESEARCH UNDERLYING THE DESIGN OF AN AUTOMATED DICTATION SYSTEM

3.1 The spoken language data
What follows is a fragment of transcription from a taped dictation session, where the speaker is dictating to a tape recorder. Pauses are indicated by [.] for a short pause (less than a second, approximately), and [-] for a longer pause.

> hi Lizzie here's some letters for you . we'll start off with . er .. Freeman and Baker paper file number 5106 that's er - Mr and Mrs Brown . write to the Inspector of Taxes at Brentford . - s er J H and Mrs S A Brown . reference 655D ...54393 --- we regret to inform you that Mrs Brown died on the 12th of October 1988 ---- er . new paragraph --- we note that we still await . er receipt of the . revised assessments for 1985 86 and 1986 87 . although we did receive a revised assessment for 1987 88
> (ETC. THEN LATER)
> - yours faithfully - . can you do me a memo on file number 1495 please Mr Gray K R Gray telephoned on . Friday . 28th of October . to indicate the building society interests that are required . er if you could knock up a memo from the details that I've scribbled down there please - now we need to write to another one - there's somebody's snuffed it - erm ---- which will be . file number 1008
> (ETC.)

3.2 Analysis
Our first step in the analysis is to identify the larger discourse structures in the data. In terms of modelling the data to reveal task structures it is evident from the extract above that the highest level of task identifiable here is that of dictation itself (as distinct from other office based activities which might be in progress). The textual evidence for this is at the beginning of the extract, which was, in fact, the beginning of the session - "hi Lizzie here's some letters for you". This can be regarded as a general initiatory sequence, equivalent perhaps, in an automated system, to calling up the appropriate software package. Trivially, it is identified by coming at the start of a text and including one of a closed set of greetings and the naming of the addressee. The word "start" in the first directive below is also, of course, a sign of the first in a series of activities.

What follows this *initiation* is a sequence of *directives* to the secretary - "we'll start off with . er .. Freeman and Baker paper file number 5106 that's er - Mr and Mrs Brown . write to". This sequence, like the initiation, is clearly not intended to be transcribed by the secretary as part of the text of the final letter, rather it is to identify which headed paper to use, and which file it is relevant to. This sequence is, then, although still not at the level of direct text for inclusion in the document, further down the task hierarchy as made explicit by the spoken natural language, in that it is specifically referring to a particular case/client/file.

At this point in the extract we encounter the first piece of material destined for inclusion in the final document - "the Inspector of Taxes". We can think of this as *text*, as it will be directly transcribed as part of the text of the finished letter. This is followed by a spoken language *macro*, which is a device frequently found in dictation data - "at Brentford". Clearly this is inadequate as a piece of typed text in a finished letter, the

secretary must flesh it out with the specific address.

Viewing the whole transcript of this dictation session we can see the same kind of structures in operation, although frequently different terms are used to signal them, particularly in the case of directives. The executive in our experiment uses different directive tokens to signal different kinds of final textual output. Letter production is signalled by "write to", and memo production is signalled by "do a memo" or "knock up a memo".

3.3 Modelling from the spoken language data

In the process of using spoken natural data as the basis for an automated system it is important to model the tasks which will make up the final system from the language model, not the other way round as is so often the case. In strongly language-based applications, it is only by constantly referring back to the language model that the designer can ensure that the system is being built to accommodate what the end users actually do. The spoken language model for the above extract can be shown as below:

Spoken language structures	Evidence from data
1. Initiation	here's some letters for you
2. Directive	we'll start off with . er ..
	Freeman and Baker paper
	file number 5106 that's er -
	Mr and Mrs Brown .
	write to the
3. Text	Inspector of Taxes
4. Macro	at Brentford

This is a first cut model of the fragment, and as the above shows, later refinements will require some of the model components to be sub-divided. In the directive section above there are five subsections which perform different functions, ranging from the very general such as "we'll start off with" and "write to", through to the very specific, such as "Mr and Mrs Brown". It is interesting to note that the pauses, filled and unfilled, tend to occur at the transition points between these subsections. This phenomenon repeats itself throughout the text. The old idea that the 'hesitations ' in speech were a sign that it was inferior in some way to printed text does not fit the facts. The pauses are every bit as important as the words and perform a similar signalling function to the stressing and intonation.

If we move now to a consideration of modelling the task of the secretary based on these spoken natural language structures, we can see how this throws some light on the subdivision of the component *directive*.

FUNDAMENTAL RESEARCH UNDERLYING THE DESIGN OF AN AUTOMATED DICTATION SYSTEM

Spoken language structures	Evidence from data	Corresponding task
1. Initiation	here's some letters for you	Start word processor
2. Directive	we'll start off with . er ..	Open new file
	Freeman and Baker paper	Get correct folder
	file number 5106 that's er -	Get correct file
	Mr and Mrs Brown .	Check file
	write to the	Type letter heading
3. Text	Inspector of Taxes	Transcribe
4. Macro	at Brentford	Find address
		Type address

4. FURTHER WORK AND FUTURE DEVELOPMENTS

As indicated in section 3 above, the next step in the analysis and modelling procedure is to subdivide the various components of the first cut model in order to identify the substructures used in the spoken natural data. This will be done on the transcribed material we have already collected, which is all from one speaker. The models produced from this will of course give us a very specific overall model of that speaker's dictation-based spoken natural language. The same procedure will be followed with similar data collected from other speakers in a range of different office environments, and the output from the analyses of the new data will be used to augment and refine the original model. At each stage, the task model will be updated to take account of the language models.

We are currently developing a prototype demonstrator for an automated dictation system, based on the spoken language models. This will be used to test the validity of our models and methodology. The first version of the prototype will be in operation by Christmas 1993.

5. REFERENCES

[1] Cheepen, C., 'Speech Aids for the Handicapped: Design and Developments at the University of Hertfordshire', this volume.

[2] Diaper, D. **Task Analysis for Human-Computer Interaction**, Chichester, Ellis Horwood Limited, 1989

[3] Young, R.M. & P. Barnard *The Use of Scenarios in Human-Computer Interaction Research: Turbocharging the Tortoise of Cumulative Science*, ACM, 1987

[4] *Literature Survey and Recommendations*, Deliverable no. 1, Research Project No. A114929/P, British Telecom and Hatfield Polytechnic, 1989

[5] Diaper, D. *Task observation for Human-Computer Interaction*, in Diaper, D. (ed) **Task Analysis for Human-Computer Interaction**, Chichester, Ellis Horwood Limited, 1989

[6] Goldkuhl, G. and K. Lyytinen *A language action view on information systems*, **Proceedings of the Third International Conference on Information Systems**, Ann Arbor, December 1982

[7] Bowers, J. & J. Churcher *Local and Global Structuring of Computer Mediated Communication: Developing Linguistic Perspectives on CSCW in COSMOS*, ACM, 1988

[8] Winter, E.O. **Towards a Contextual Grammar of English**, George Allen & Unwin Ltd., 1982

[9] Halliday, M.A.K. **Intonation and Grammar in British English**, Mouton, 1968

[10] Monaghan, J. *The Pragmatics of Human-Computer Interaction - an overview*, International Pragmatics Conference,1990

[11] *Intonation in Computer Generated Dialogue*, Alvey Project MMI/SP no. 001

[12] Schegloff, E. *Sequencing in conversational openings*, **American Anthropologist**, 70, No. 4 pp 1075-95 1968

[13] Schegloff, E.A. & H. Sacks *Opening up Closings*, in Turner, R. (ed) **Ethnomethodology**, Harmondsworth, Middx., Penguin Books, 1974

[14] Monaghan, J. *On the Signalling of Complete Thoughts*, in Benson, J.D. & W.S. Greaves (eds) **Systemic Perspectives on Discourse Vol. 1**, Ablex, Norwood, N.J., USA, 1985, pp. 373-82

[15] Cheepen, C. **The Predictability of Informal Conversation**, Pinter Publishers Ltd., 1988

[16] Cheepen, C. & J. Monaghan **Spoken English: A Practical Guide**, Pinter Publishers Ltd., 1990

[17] Williams, B. & D. McKelvie *A Preliminary Statistical Analysis of Contextually-Conditioned Segmental Durations Using Automatically-Segmented Data*, **The Collection and Measurement of Spoken Language**, Institute of Acoustics Speech Group Meeting, 23 February 1990