

An Ethnography of Speech Recognition

Ben Kraal¹, Penny Collings¹, Anni Dugdale², Michael Wagner¹

¹School of Information Sciences and Engineering

²School of Business and Government

University of Canberra

Email: {ben.kraal, penny.collings, anni.dugdale, michael.wagner}@canberra.edu.au

Abstract (Heading – abstract)

This short paper reports the preliminary findings of our ethnographic investigation into the use of commercial speech recognition technologies in real work places. We discuss how our ethnographies have shown that the usability of speech recognition is equally about the technology and the social environment in which it is used. We also discuss our method and its wider implications to areas other than speech recognition.

Keywords

Ethnography, speech recognition, actor-network theory, locales framework, ubiquitous computing.

INTRODUCTION

In this short paper we describe the preliminary findings of our ethnographic study into the use of commercial, off the shelf, speech recognition technology. The paper briefly describes the method used and presents a short ethnography of the situations we investigated. Finally, we describe our findings on the use of speech recognition technology and what we believe are the implications for other “intelligent” software and ubiquitous computing.

Ethnographic techniques are increasingly used in studies of technology. Suchman (1987) used ethnographic techniques to great effect, demonstrating the poverty of man-machine interaction with photocopiers as well as critiquing the wider field of Artificial Intelligence. Many other studies have since used ethnographic techniques to study diverse technologies (Heath and Luff, 2000). As computer scientists use ethnography as an input to a design process some (Button and Dourish, 1996; Hemmings and Crabtree 2002) have begun to redefine ethnography, as it is used in computer science, to make it more usable to computer scientists. We take our cue from these works and see our study as following in their footsteps.

Though this paper describes our ethnography of speech recognition users, it is possible to extend our method to other areas such as ubiquitous computing (Ubicomp) and “intelligent software” (eg Agents and other recognition based technologies). Ubicomp is a new area of interest for many computer scientists though how people will begin to use ubiquitous technology that is more advanced than that which is currently available is poorly understood (if indeed that can be known at all). Our application of ethnographic methods could be extended to ubiquitous computing situations to reveal the non-technological aspects in the use of the technology. Regarding intelligent software, it is common for proponents to focus on quantitative measures of performance such as recognition rate to evaluate the usability of their software. While these measures are valuable, our preliminary findings here show that the usability of speech technology is more complex than quantitative measures can reveal and that an appreciation of the social aspect of the technology as it is used is important if the software is going to be truly usable in a real work (or play) situation.

BACKGROUND TO THIS PAPER

We were approached by the Chief Magistrate of the ACT Magistrates Court (the Court) to investigate the introduction of speech recognition technology to the courtroom for use by the magistrate in the process of determining outcomes. Determining outcomes, as we have come to understand it, is a highly charged moment in the Court when the magistrate speaks an outcome for the case that he or she is hearing. An outcome may be a sentence, for example a fine or jail term or it may be the decision to set a case over to allow all the parties to the case more time to gather relevant information. An outcome may also be a procedural decision specific to the Court such as a request by the magistrate for any number of specialised reports that are used to inform the actual sentence when it is finally delivered. The question we were attempting to answer was: what form would speech recognition technology take at the Court?

After some preliminary ethnographic work at the Court it emerged that the magistrate’s act of speaking an outcome was not an event that was self contained but the beginning of a process distributed in space and time throughout the Court. We began a deeper ethnographic investigation of the processes involved in determining and recording outcomes of cases. We began to understand that the process involved many different Court workers, each

performing detailed work that contributed to the final outcome. It would be usual to only perform one ethnographic study before beginning design work, however, the more we began to understand about how the Court worked, the more it began to emerge that any design for speech technology at the Court would involve not only software design, but also work process design.

Designing new work practices to suit a technology as poorly understood as speech recognition is fraught with difficulty, so we began two ethnographies of work places where speech recognition software had already been introduced to try to understand the changes that speech recognition makes on a work place and work practice and what, if any, changes the work practice makes to the software. These ethnographies revealed that the introduction of speech recognition technology to individual users had effects throughout the whole business, even to users who did not work directly with the speech recognition users.

There are two workplaces we saw speech recognition software in use: the first was the Hansard Section at Parliament House and the second was composed of various offices within the public service in general. Since the software used in both locations is basically the same off-the-shelf software, the different perceptions of its use are interesting and revealed to us a lot about what it actually takes to make speech recognition software usable in a productive environment. In the next sections the reasons for the use of speech recognition technology in each workplace are explained.

The Hansard Section at Parliament House

At Parliament house, the Hansard Section is concerned with producing the document called “Hansard” that is a record of what was spoken in parliament. The current work practice has some Hansard editors using commercial speech recognition software to re-speak the words of the parliamentarians to transform their speech into text that can be edited and formatted into the document called Hansard.

The editors have a great deal of freedom in how they use the speech recognition software. They may only use it for re-speaking and do all of their editing with mouse and keyboard, or they may do some editing by hand and some by voice, or they may do all of their editing by voice. Within the Hansard section, speech recognition software is treated as another business tool, in much the same way as laser printers, monitors or spreadsheet software is treated in other businesses or government departments.

General Public Service Staff with Soft Tissue Injuries

Within the public service there are many office workers who begin to have pain in their hands and arms when typing or using a mouse. A collective term for injuries of this type is “soft tissue injuries” though they are commonly known as RSI (Repetitive Strain Injury) which is actually a specific kind of soft tissue injury. The work of each person observed was different though all interviewees performed high volume typing.

When the pain of their injury prevents them from working, a worker’s doctor or physiotherapist may recommend that they use speech recognition software to prevent the injury from re-occurring. Depending on the injured person’s position in their organisation, this may be very easy to achieve or it may be difficult. Often, negotiating an injured person’s return to work with speech recognition software involves the interaction of the IT department, the Occupational Health and Safety department, the injured person, their doctor or doctors, their manager and possibly even more parties.

METHOD

The method that evolved to answer our original question, then, is to do an ethnography for a “greenfield” location, one where a new technology is not yet in use, and to also do an ethnography for several situations where the technology, or a variant of it, is already in use. The ethnography of the greenfield site reveals its essential properties and the ethnographies of the in-use sites, when analysed together, can allow designers to imagine potential outcomes in the greenfield location and to design interfaces or work practices to mitigate undesirable outcomes. In this paper, we report on the ethnography of the situations where we examined speech technology in use.

The general Public Service speech technology users were self-selected by placing a call for volunteers in a newsletter produced by a local speech recognition trainer. The Hansard users were selected by asking the manager of the Hansard section who he would recommend to be interviewed.

In some fields, ethnographies stand alone as descriptions of the detail of lived experience however in the fields of Information Systems, Information Technology or HCI, a choice must be made on how to analyse an ethnography.

In this case we have chosen to use two different lenses with which to view the ethnographies. Those lenses are Actor-Network Theory (Law, 1992) and the Locales Framework (Fitzpatrick, 2003).

Actor Network Theory (ANT) is a misnamed (in that it is not actually a theory) way of looking at situations to discover the power structures, how animate and inanimate "actors" relate and the structures that are set up to allow a particular situation to continue existing. Using the lens of ANT has shown the importance of enrolling the entire organisation in the new network established by the technology. ANT uses the word enrolling to mean something akin to "including in the process" and the term "network" does not mean a network of data carrying wires but one of people, places and objects.

The Locales Framework is a comprehensive method for examining CSCW (and CSCW-like) situations where many people work co-operatively using software. The Locales Framework can reveal the relationship between the social and technical. In this paper we use the framework very simply to emphasise the importance of "place", the environment in which the software is used.

PRELIMINARY FINDINGS

Our preliminary findings, while not revolutionary, can be said to show that the use and usefulness of speech recognition technology is not only a property of the software itself but is also dependant on the willingness of the organisation to adapt to the demands of the software and the physical work environment in which the software is used.

The importance of enrolling the organisation in the technology

In the Hansard Section of Parliament House, the editors are given training in the software and high quality microphones, earphones and computers than can cope with the rigors of using speech recognition. The IT support staff have helped the editors create voice macros to automate some of the repetitive formatting tasks that are part of creating the Hansard document. This is in contrast to the experience of the general public service users who are often not provided any training, other than that which the software manual provides, and have to petition their managers for access to computers powerful enough to cope with the system load of speech recognition software and everyday applications working together.

In the Hansard section, speech recognition software is an accepted part of the work practice involved in creating the Hansard document. It is one of many tools that the editors have at their disposal and some editors do not use the speech recognition software, preferring to use other transcription methods such as CAT, Computer Aided Transcription, which is a chording-keyboard method. In the general public service, speech recognition is outside of the work practice and anyone who uses it is seen as disruptive, if not by their peers, then by some or all of the IT support staff, the Occupational Health and Safety staff or their managers. The speech recognition users interviewed reported that they had difficulty in getting IT staff to deal with their computer problems because the IT staff would blame all problems on the speech recognition software which was effectively out of their scope for support.

The public service staff were effectively alone in their organisations and felt they were treated as "problem users" by their department at large. They were left to solve their own problems with the software despite it being in their department's interest for them to remain productive. In the Hansard section, the reverse is true: the department has adopted speech recognition software as part of the general work practice because it offers specific advantages. Typical learning times are 6 months to attain proficiency with the speech recognition software and more than three years for CAT.

The work environment matters

The work environment, the social world in which technology is used, changes the users' experience of the technology. In the Hansard section, the work environment of the editors has been modified to be more accepting of the special needs of speech recognition. There are two types of offices that the editors work in. The first is a long thin room that is an office for one person. The room has a window, a computer and the speech recognition hardware as well as the other paraphernalia that the editors use to do their job (dictionaries, lists of Australian place names, lists of sitting members' names and electorates etc). The room has a door that can be shut to prevent the sound of using the speech recognition software from intruding into other people's work environment. The other offices are shared. We were surprised to see speech recognition users working in shared offices as it is generally accepted that other speech is one of the most confounding factors to good recognition accuracy. The shared offices are large square rooms with four editors working in them. Each editor's desk faces into a corner of the room. There is sound deadening material along the walls directly in front of each editor, behind their computer.

The general public service speech recognition users we spoke to, and whose office environments we observed, are more variable in their conformance with good speech recognition practices. Only one of those interviewed had their own office; the others worked in open-plan or cubicle offices often with non-speech recognition users working close by. Some users had taken the step of asking to be relocated away from other people, partly to minimise the recognition errors cause by the normal work noise of other people but also partly because the users

felt that using speech recognition software was disruptive or annoying to their co-workers. This relocation had the effect of isolating the speech recognition users from their co-workers.

REFLECTIONS ON THE METHOD

This method of investigating two case studies which can be contrasted with each other has proved useful in revealing the properties of the greenfield situation that must be respected and the properties of the technology-in-use cases that contribute to making the software usable.

Our method of case and counter cases could also be applied in wider fields where technology that is in its infancy is being rapidly developed and where the use of the technology cannot be adequately captured in a laboratory. For example, ubiquitous computing is only just beginning but by studying how the current level of development is used by people in their daily lives it may be possible to develop more advanced technologies that are more easily appropriated by users in their work and play.

We believe that our findings are complementary to the need for software advances in speech recognition and other kinds of "intelligent software". In speech technology literature the social is often ignored in favour of the technical (for example: Deng and Huang, 2004). Our ethnographies have shown that the social aspects of this technology must be regarded with equal importance if the technology is to move from tightly controlled situations such as telephone response systems into real work situations that are significantly more complex. Similarly, speech recognition vendors often state that current off-the-shelf technology is easy to use but this fails to address the issues of how the technology fits into the wider workplace with respect to established work practices, the additional burden on technical support staff, and even the physical layout of offices and cubicles.

Speech recognition software is not seen as particularly social software but our ethnographies show that it has the potential to make significant changes to the social nature of a workplace. Software that is explicitly social in nature, such as ubicomp, will also cause changes in the social structure of workplaces, and other social places, that are not yet understood. A method such as that described here could allow such potential changes to be anticipated before they occur. Additionally, our method could be used to investigate and design new technologies in ways that allow them to be more easily integrated into existing social situations. As technologies move off the desktop and into the world, the social aspects of technologies will be increasingly important.

REFERENCES

- Button, G and Dourish P. (1996) On "Technomethodology": Foundational Relationships between ethnomethodology and system design. *Human Computer Interaction*, n4, v13, 395-432.
- Deng, L and Huang, X. (2004) Challenges in Adopting Speech Recognition. *Communications of the ACM*, v47, 1, 69-75.
- Fitzpatrick, G. (2003) *The locales framework : understanding and designing for wicked problems*. Kluwer Academic Publishers, Boston.
- Heath, C and Luff, P. (2000) *Technology in Action*. Cambridge University Press, New York.
- Hemmings, T and Crabtree A. (2002) Ethnography for Design? *Proceedings of the International Workshop on "Interpretive" Approaches to Information Systems and Computing Research* 122-124
- Law, J. (1992). "Notes on the Theory of the Actor-Network: Ordering, Strategy and Heterogeneity." *Systems Practice*, 5,379-393.
- Suchman, L. (1987) *Plans and Situated Action*, Cambridge University Press.

ACKNOWLEDGEMENTS

The authors would like to thank the ACT Chief Magistrate, the interviewees at Parliament House, the employees of the ACT Magistrates Court and the anonymous Public Service employees for their kind cooperation.

COPYRIGHT

Ben Kraal, Penny Collings, Anni Dugdale, Michael Wagner © 2004. The authors assign to OZCHI and educational and non-profit institutions a non-exclusive licence to use this document for personal use and in courses of instruction provided that the article is used in full and this copyright statement is reproduced. The authors also grant a non-exclusive licence to OZCHI to publish this document in full in the Conference Papers and Proceedings. Those documents may be published on the World Wide Web, CD-ROM, in printed form, and on mirror sites on the World Wide Web. Any other usage is prohibited without the express permission of the authors.