The TalkBank Project

Brian MacWhinney

1 Introduction

Recent years have seen a phenomenal growth in computer power and connectivity. The computer on the desktop of the average academic researcher now has the power of room-size supercomputers of the 1980s. Using the internet, we can connect in seconds to the other side of the world and transfer huge amounts of text, programs, audio and video. Our computers are equipped with programs that allow us to view, link and modify this material without even having to think about programming. Nearly all of the major journals are now available in electronic form and the very nature of journals and publication is undergoing radical change.

These new trends have led to dramatic advances in the methodology of science and engineering. However, the social and behavioural sciences have not shared fully in these advances. In large part, this is because the data used in the social sciences are not well-structured patterns of DNA sequences or atomic collisions in supercolliders. Much of our data is based on the messy, ill-structured behaviours of humans as they participate in social interactions. Categorizing and coding these behaviours is an enormous task in itself. Moving on to the next step of constructing a comprehensive database of human interactions in multimedia format is a goal that few of us have even dared to consider. However, recent innovations in internet and database technology provide excellent methods for building this new facility. Unlike the structured databases of relational database programs like Excel or Access, the new database formats are designed specifically to handle messy, ill-structured data, such as that found in human communication. XML tools developed by the World Wide Web Consortium or
W3C (http://w3c.org) can be applied to represent language data. The interlocking framework of XML programs and protocols allows us to build new systems for accessing and sharing spoken language data. At the same time, improvements in computer speed, disk storage, removable storage and connectivity are making it easier and easier for users with only a modest investment in equipment to share in this revolution.

Among the many fields studying human communication, there are two that have already begun to make use of these new opportunities. One of these fields is the child language acquisition community. Beginning in 1984, with help from the MacArthur Foundation, and later National Institute of Health and National Science Foundation, MacWhinney and Snow (1985) developed a system for sharing language-learning data called the Child Language Data Exchange System (CHILDES). This system has been used extensively and forms the backbone of much of the research in child language of the last 15 years.

A second field in which data-sharing has become the norm is the area of speech technology. There, with support from Defense Advanced Research Projects Agency (DARPA) and a consortium of businesses and universities, Mark Liberman and Steven Bird have organized the Linguistic Data Consortium (LDC). The corpora of the LDC now also function as the backbone for the development and evaluation of technologies for automatic speech recognition and generation.

Recognizing the positive role of data-sharing in these two fields, and the need for improvement in infrastructure for the social sciences (http://vis.sdsc.edu/sbe/), the National Science Foundation provided funding for a new project called TalkBank (http://talkbank.org). The goal of the project is to support data-sharing and direct, community-wide access to naturalistic recordings and transcripts of human and animal communication. TalkBank has identified these seven shared needs:

1. Guidelines for ethical sharing of data
2. Metadata and infrastructure for identifying available data
3. Common, well-specified formats for text, audio and video
4. Tools for time-aligned transcription and annotation
5. A common interchange format for annotations
6. Network-based infrastructure to support efficient (real time) collaboration
7. Education of researchers to the existence of shared data, tools, standards and best practices.