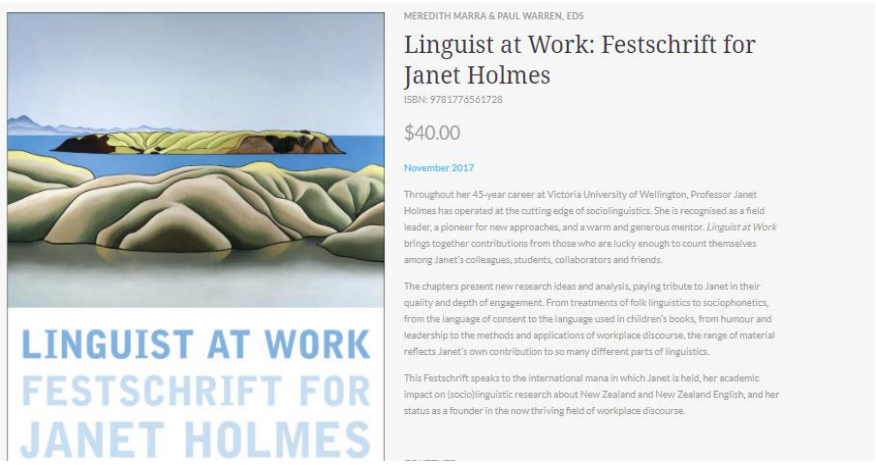
Please cite as:

Stubbe, Maria 2017. Evolution by Design: Building a New Zealand Corpus of Health Interactions. In Marra, M. & Warren, P. (Eds), *Linguist at Work: Festschrift for Janet Holmes*. Wellington: Victoria University Press, pp 196-214.

ISBN: 9781776561728

Publication date: November 2017



MEREDITH MARRA & PAUL WARREN, EDS

Linguist at Work: Festschrift for Janet Holmes

ISBN: 9781776561728

$40.00

November 2017

Throughout her 45-year career at Victoria University of Wellington, Professor Janet Holmes has operated at the cutting edge of sociolinguistics. She is recognised as a field leader, a pioneer for new approaches, and a warm and generous mentor. *Linguist at Work* brings together contributions from those who are lucky enough to count themselves among Janet's colleagues, students, collaborators and friends.

The chapters present new research ideas and analysis, paying tribute to Janet in their quality and depth of engagement. From treatments of folk linguistics to sociophonetics, from the language of consent to the language used in children's books, from humour and leadership to the methods and applications of workplace discourse, the range of material reflects Janet's own contribution to so many different parts of linguistics.

This Festschrift speaks to the international mana in which Janet is held, her academic impact on (socio)linguistic research about New Zealand and New Zealand English, and her status as a founder in the now thriving field of workplace discourse.

http://vup.victoria.ac.nz/linguist-at-work-festschrift-for-janet-holmes/

11. EVOLUTION BY DESIGN: BUILDING A NEW ZEALAND CORPUS OF HEALTH INTERACTIONS[1]

Maria Stubbe

1. Introduction

There is a large volume of published research on language and communication in healthcare, but for a long time this was very much a game of two halves: on the one hand, studies of patient–professional communication from an 'insider' professional perspective within the clinical literature; and on the other, research into more nuanced aspects of directly observed interactions from within humanities and social science disciplines such as linguistics, discourse analysis, conversation analysis and sociology (Sarangi 2004). Historically, clinical studies have also tended to rely heavily on reported data from questionnaires and surveys, or on high-level coding of consultation structure and content, with the aim of answering questions about how communication influences health outcomes. By comparison, studies focusing in close detail on the discursive and linguistic features of health interactions were most often undertaken by researchers from outside the healthcare world, and seldom explicitly addressed matters of application and practical relevance.

More recently this gap has started to close. There has been a marked increase in interdisciplinary applied work on communication in healthcare settings, and researchers across the spectrum are also

starting to recognise how specialised methodological tools from sociolinguistics and conversation analysis might be harnessed, in conjunction with technical innovations in software development and digital recording, to analyse large scale health-related multimodal corpora (Crawford, Brown and Harvey 2014). However, by comparison with other linguistic corpora, the construction of health-related collections raises a special set of ethical, methodological and governance questions. This chapter outlines the key issues likely to face developers of such collections, and discusses the building of a New Zealand corpus of health interactions to provide a practical illustration.

2. Background

Clinical consultations are one of the more challenging communicative settings to research because they typically involve a patient and health professional talking confidentially about highly personal matters in a small room, making direct observation problematic. Video recordings without a researcher present offer a unique window into the 'black box' of actual health encounters. They allow researchers to investigate in detail how patients and providers interact with one another in real-life situations, and to tease out specific practices that foster or hinder effective communication and improved outcomes. Indeed, the value of capturing and analysing audio-visual consultation data has been recognised by clinical educators since the early days of portable recording technology. An influential example of this is the foundational study of clinical decision making, *Doctors Talking to Patients* (Byrne and Long 1976). Byrne and Long audio-recorded over 2000 consultations with 71 British GPs, using a coding scheme to describe in broad terms how doctors communicated in each interaction. A more recent and refined development of this approach is the Roter Interaction Analysis System (RIAS), which remains very influential in mainstream clinical communication research and education (e.g. Roter and Larson 2001).

In the last two decades, research using the micro-analytic tools of interactional sociolinguistics and conversation analysis has gained increasing traction in health communication research. This interdisciplinary body of work has generated important theoretical

and applied insights into the fine-grained interactional practices and structures typical of encounters between patients and health professionals (e.g. Heritage and Maynard 2006; Hudak, Clark, and Raymond 2012; Robinson 2012). More recent research has added a focus on multimodal aspects of interaction and the influence of health technologies and medical informatics on communication in a range of different healthcare settings (e.g. Dowell et al. 2013; Mondada 2016; Swinglehurst et al. 2014). The same technical advances have enabled focused video-ethnographic fieldwork in specific healthcare settings such as intensive care units (e.g. Carroll, Iedema and Kerridge 2008; Wyer et al. 2017). A related strand of work involves the gathering of high quality audio and video recordings of in-depth narrative interviews to explore patient experiences of health and illness, with the dual purposes of sociological or sociolinguistic research and the online publication of curated excerpts for educational purposes (e.g. Pope and Davis 2011; Ziebland and Herxheimer 2008).

Collecting authentic recordings of health interactions and health-related conversations is ethically and logistically very challenging (Parry et al. 2016), and data collection and processing are invariably time-consuming and resource-intensive. Going through the additional technical and ethical hoops to develop a permanent archive is certainly possible, but is more demanding still, and requires considerable foresight and perseverance, as well as overcoming the major barrier of resourcing (Jepson et al. 2017). Moreover, there are particular ethical and medico-legal constraints around the use and re-use of health information that prevent the creation of formal and freely accessible health-related corpora along the same lines as other large international electronic corpora of spoken language. As a consequence, studies of health interaction are often based on analysis of relatively small, 'one-off', purposively collected sets of consultations, or occasionally on high-level content analysis of larger routine data sets, as was the case with the Byrne and Long (1976) study, neither of which may ever be used again beyond the life of a particular project.

More recently, despite the challenges, some health communication research groups have been starting to generate substantial and useful collections in the course of their own work, with permissions in

place to archive the data and make it available to others. Table 1 provides summary information about several electronic archives of English language health interactions or other health-related data from the UK, USA and New Zealand that exist currently, each with somewhat different origins and objectives. These collections include video and audio recordings of health interactions of various kinds, ethnographic or narrative interviews, and health-related electronic texts. They are all restricted-access multimedia corpora based at universities or medical schools, with curated data available to authorised researchers under a variety of access and governance arrangements that range from local oversight by the originating research unit through to institutionally managed archives. Such collections offer cost-effective opportunities for secondary analysis of health communication and related data to answer a wide range of research questions, using a variety of methodologies. They also allow researchers to access substantially larger and/or more varied collections than would be possible if collecting data *de novo*, to use existing data as a test bed for designing another more targeted study, or to engage in comparative or collaborative research, while at the same time reducing the burden on participant groups.

However, such collections of health-related data are still the exception. Most often researchers are working in isolation, with legal, ethical or practical constraints making it impossible to share health interaction data outside the original research team; in some cases (e.g. in parts of the USA), video or audio recording of consultations is not readily allowed at all due to medico-legal sensitivities (e.g. Barone 2012). Moreover, in many countries regulatory or ethical conditions still commonly require researchers to destroy confidential health information once a specific project has ended. Even where permission for re-use has been granted by participants and is ethically approved, sharing sensitive health-related data safely and confidentially will always be somewhat problematic, and ethics committees will require stringent access and governance protocols to be in place to protect participants' privacy and confidentiality, especially as it is not practical to comprehensively anonymise audiovisual material (Parry et al. 2016). From the user's perspective, data quality in informal or ad hoc collections that can be shared is often variable, possibly with

Table 1: Selected health-related English language corpora

One in a Million Primary Care Consultations Archive A restricted access database managed by Bristol University. It comprises 300 video-recorded GP consultations along with verbatim transcripts and a range of associated data including demographics, consultation records and standardised questionnaires. The consultations were recorded in 12 General Practices in the West of England in 2014–2015, with permission from participants for future re-use in research and education (Jepson et al. 2017). *www.bristol.ac.uk/primaryhealthcare/researchthemes/one-in-a-million/*
Carolinas Conversations Collection A restricted access digital collection of transcribed audio and video recordings of conversations about health held at the Medical University of South Carolina Library. It has two cohorts: 125 unimpaired multi-ethnic older speakers with a chronic condition, and a longitudinal set of 400 conversations with 125 persons with dementia. It includes information about health literacy, health status, and cognitive function (Pope and Davis 2011). *http://carolinaconversations.musc.edu/about/collection*
Nottingham Health Communication Corpus 500,000 words of health language data in a range of different settings including a range of practitioner–patient exchanges and patient/service-user narratives. The NHCC is maintained by an interdisciplinary health language research group (HLRG) at Nottingham University, and incorporates multiple data sources, from both computer-mediated communication and from spoken texts (Crawford, Brown and Harvey 2014). *http://www.nottingham.ac.uk/research/groups/hlrg/index.aspx*
Health Experiences Research Group (HERG) Archive Contains over 3,000 interviews with patients, carers and other family members comprising full sets of topic-based interviews (35–50 per collection) and supporting documents. Interviews in the archive are copyrighted to the University of Oxford and are available, under licence, to approved qualitative researchers for secondary analysis (Zieband and Herxheimer 2008). *http://www.healthtalk.org/research/use-our-data*
ARCH Corpus of Health Interactions The ARCH Corpus is a restricted access collection of approximately 500 digitised video/audio-recorded health encounters and 250 interviews, along with related ethnographic and demographic data collected in New Zealand since 2003. This material is permanently archived at the University of Otago for approved future use in research and education (Applied Research on Communication in Health Group 2017). *http://www.otago.ac.nz/wellington/research/arch/corpus/*

little in the way of ethnographic or demographic information, input from clinicians or longitudinal data to enrich the analysis and assist interpretation, and with limited search functionality (Barnes 2016; Barone 2012).

When designing a project involving audio-visual health interaction data, it therefore makes sense, where possible, to proactively maximise the potential for ongoing systematic uses of the data in addition to all the usual ethical and methodological requirements, even if the immediate intention is not to set up a formal corpus or archive. The remainder of this chapter provides an overview of how key issues of ethics, practical methodology and governance have been addressed during the evolving process of building the ARCH Corpus of Health Interactions at the University of Otago in New Zealand.

3. Building the ARCH Corpus of Health Interactions

Members of the interdisciplinary Applied Research on Communication in Health (ARCH) Group at the University of Otago, Wellington, started collecting and analysing video recordings of naturally occurring interactions between health practitioners and patients in 2003, as part of a multidisciplinary project on clinical decision-making in referrals for elective surgery in New Zealand. The team at that time included two clinicians (a senior GP academic and a nurse researcher with training in discourse analysis), a health services researcher, a medical sociologist and an interactional sociolinguist, four of whom have continued as co-directors of the ARCH Group to the present day. The ARCH Corpus of Health Interactions was formally established in 2005 and has continued to grow and evolve since then. Its location at a trusted university medical school and the complementary professional networks and expertise of the core team members have been crucial factors in gaining ongoing access to research sites in the health sector, and in the creation of a robust, multi-purpose technical platform.

The ARCH Corpus has now developed into a searchable digitised collection of New Zealand healthcare interactions and related data which are permanently archived (with consent of participants) for use in research and education. A key aim has been to maximise its relevance and usability for a wide range of potential end-user

groups, including health professionals, clinical educators and patient groups, as well as academic researchers and students from a variety of disciplines. The development of a formal collection of video-recorded health interactions of this type represented a significant innovation internationally when the ARCH Corpus project was first initiated, and it remains unique in New Zealand.

3.1 Structure and composition

The ARCH Corpus is not formally structured like a linguistic corpus; rather it is a 'living data bank' that comprises digitally recorded health interactions, research interviews and associated information collected progressively in the course of successive research projects (see Table 2).

Table 2: ARCH Corpus data sources*

Studies involving primary data collection	**Data collection**
Interaction Study (IS)	2003–2005
Surgeons Study (SS)	2006
Tracking Study (TS)	2006–2009
Diabetes Tracking Study (DS)	2009–2012
Demystifying Addiction (DA) (interviews)	2012
Talking About Obesity and Overweight (TAbOO)	2012–2013
Interpreters Study (IN)	2012–2013
Health Navigators and Interpreting (HNI) (interviews)	2015
Ante-natal Clinic Study (AN)	2016
Diabetes Stories (interviews)	2016–2017

*Further details are available on the ARCH website (Applied Research on Communication in Health Group 2017)

Integral to the structure and maintenance of this corpus is a comprehensive data management system. This includes a searchable relational database which also functions as a detailed catalogue of linked datasets, with hyperlinks to a separate file directory

containing audio and video recordings, standardised logs and field notes, transcripts, medical notes, and other associated information. This collection and its custom-designed data management system were specifically designed to evolve over time and to accommodate the ethical, logistical and methodological challenges inherent in collecting and analysing sensitive personal and professional data from a variety of healthcare settings.

Table 3: Composition of the ARCH Corpus (as at July 2017)

<table>
<tr><th colspan="2">Primary audio-visual data</th></tr>
<tr><td>478 consultations/health interactions
156 related research interviews
Involving:
533 participants:
Patients / GPs, nurses, surgeons, specialists, allied health professionals / other participants e.g. interpreters, family
38 health care sites
General practices, community clinics, hospital settings</td><td rowspan="2">Audio and video recordings
Content summaries
Timed content/action logs
Base transcripts (orthographic plus key interactional features)
Selected CA transcripts
Verbatim transcripts (interviews)
Derived clips and extracts
Case inventories (longitudinal data)</td></tr>
<tr><td>77 ‘standalone’ research interviews & focus groups</td></tr>
<tr><th colspan="2">Other ethnographic & contextual data</th></tr>
<tr><td colspan="2">Participant information (Demographics & background information, ‘on the day’ consultation notes, referral letters)
Field notes (Site descriptions, incidental observations, photos, informal ‘debriefs’)
Documents (Clinical protocols/templates/guidelines, patient education materials)</td></tr>
<tr><th colspan="2">Administrative data</th></tr>
<tr><td colspan="2">Consent details (per project & per individual participant)
Data set information (linked files & information)
Data processing history (dates, formats, versions)</td></tr>
</table>

Table 3 provides an overview of the current composition of the ARCH Corpus. Data acquisition has focused on the twin aims of (i) obtaining rich interlinked qualitative data sets that allow ‘thick description’ of the discursive construction of a single interaction or a longitudinally tracked case or episode of care; and (ii) achieving

sufficient diversity to facilitate exploration of a range of clinical, interactional or linguistic questions across the corpus. The data is suitable for both micro-level analysis of individual interactions, and more macro-level ethnographic or clinically focused descriptions of communicative patterns and systems.

3.2 Motivation for setting up the ARCH Corpus

The initial impetus for setting up the ARCH Corpus came from a collective desire on the part of the core research team to continue working with the 75 hard-won video-recorded consultations from our first project. We had barely scratched the surface of what this first tranche of data might yield by the time the project funding came to an end after two years, and there was clearly tremendous scope not only for further research, but also to make use of this material for evidence-based training and professional development in clinical communication. We were also very conscious of the altruistic motivations of the research participants and the trust they had placed in us to make the best possible use of the data they contributed.

Another important influence was my previous involvement in the creation of sociolinguistic corpora, including the Wellington Corpus of Spoken New Zealand English (part of the International Corpus of English) under the direction of Janet Holmes in the 1990s. My subsequent work with Janet in establishing the Wellington Language in the Workplace Project (LWP), and helping build and organise the LWP's large collection of discourse data, had cemented in my mind the value of corpora as a way of preserving and cataloguing interactional data for future use, and the importance of engaging proactively and constructively with research participants and end users (see also Vine and Marra, this volume). It seemed a logical next step to adapt and extend the relevant principles and methods from this previous work to the setting of health communication research in order to develop a permanent collection of health interaction data that could be made available to other researchers and educators. A substantial grant awarded in 2005 for a new ARCH Group study made it possible to put this collective vision into practice by supporting the recording

of another 200 consultations and the initial design of our corpus data management system.

3.3 'Kaupapa' and ethical considerations

The Mäori word *kaupapa* can be roughly glossed as meaning 'the guiding philosophy and principles that inform a group's customary ways of doing things' (Royal 2017). The underlying kaupapa or ethos of the ARCH Group emphasises the central importance of trust and respect, both in our relationships with our research participants and stakeholders, and in the way we manage the data they have gifted to us. These ethical principles are reflected, firstly, in the close attention we pay to how we go about collecting data, and how it is stored, analysed, interpreted and used; and secondly, in the constructive 'appreciative inquiry' approach (Dick 2004) that we apply at all stages of the research process, especially in the dissemination of findings. The wider aim is to engage in 'research for and with' rather than 'research on' health professionals, patients and service users by actively maintaining a two-way dialogue, and contributing to the shared practical goal of improving health.

Health research ethics committees in New Zealand, as elsewhere, quite rightly uphold stringent standards for observational research involving personal health information. Matters that must be appropriately addressed include: obtaining informed consent; participants' rights over how their data is used; avoiding coercion and exploitation; ensuring data privacy, confidentiality and security; and restricting access to or dissemination of identifiable (non-anonymised) or potentially re-identifiable personal information. These issues are considered within an overall bioethical framework of avoiding harm, minimising the burden on participants, ensuring scientific validity, and considering the likely benefits of the research, as well as various legal and regulatory protections which must be adhered to in research and in clinical practice such as the Privacy Act 1993, the Health Information Privacy Code 1994 and the Code of Patient Rights in New Zealand (National Ethics Advisory Committee 2012).

The need to satisfy this complex set of ethical and medico-legal requirements has been a fundamental consideration in designing

the structure and processes for our corpus as a whole. We also have to carefully think through the ethical and methodological implications of archiving data for unspecified future use in addition to meeting the primary and more immediate research objectives of a given study. This adds several layers of complexity to the research design and ethical review processes for individual projects. Furthermore, during the early years in particular, it was not always easy to persuade ethics committees that video-recording health consultations was in fact an acceptable methodology, or that there were good reasons for retaining such material beyond the life of a specified project; both were unusual in health research at that time. It was thus essential to demonstrate up front that we had carefully considered protocols in place, and that we had consulted within our health professional networks and at individual research sites to ensure that we met medico-legal requirements (e.g. patient confidentiality), would not unduly disrupt 'practice as usual', and could address perceived reputational risks to individual practitioners or health professional groups. Of course these latter points were also crucial to gaining the trust and willing participation of health providers and practitioners in our research. To ensure that the structure of the ARCH Corpus and its governance processes would consistently align with our overall kaupapa, we established a set of clear methodological design principles at the outset, and these continue to guide the ongoing development of the corpus.

3.4 Methodological and technical design

The term 'evolution by design' captures the essence of our approach to constructing the ARCH Corpus. From a practical perspective, we knew from the start that this collection would have to be 'grown' organically as an additional output from successive research projects, as there was no funding to be had purely to create a data resource. We therefore invested considerable time and energy into establishing robust methodological and technical protocols for every stage of the research journey, from participant recruitment and consent through to the collection, processing, archiving and management of data. These were designed to function as a well-integrated system, broadly consistent across different ARCH Group projects, and in accord with

our kaupapa and ethical responsibilities. At the same time, we kept our system design flexible enough to allow for future technological developments and the variations that would inevitably arise as we added new projects. We also had to accommodate a diverse range of disciplinary interests and analytic approaches (including interactional sociolinguistics, conversation analysis, medical sociology, and academic primary care), and a range of theoretical, educational and clinical objectives. This too had implications for the kinds of data and metadata we needed to collect, and the formats in which it should ideally be recorded and stored.

It is beyond the scope of this chapter to go into much technical detail, but as noted earlier, a key aspect of our corpus design was the creation at an early stage of a comprehensive data management and file storage system. We used a customised version of Microsoft Access to create a multi-purpose relational database. Individual metadata records are hyperlinked to a carefully constructed document folder system housing the actual data files; these are stored outside the database programme itself on the same secure university network. The database facilitates the efficient processing, archiving, linking and retrieval of each project's multi-media and associated data files, as well as storing de-identified demographic data and administrative information such as consents granted and conditions of use for data subsets or individual items. It also functions as a user-friendly portal and search tool for use in sample selection and data analysis, and as a permanent information management system for the ARCH Corpus as a whole. Audio-visual data files and full transcripts are securely encrypted if they need to be accessed offsite, and the whole corpus is protected by multiple backups. This hybrid system works well for a heterogeneous and open-ended collection of interlinked data sets such as the ARCH Corpus.

The ARCH Corpus design also encompasses a set of standardised processes and documentation for data collection and processing, which integrate seamlessly with the database and archiving system described above. Free-text field observations, site information and technical data relating to data collection are routinely recorded onto templates. Research participants fill out a standard demographic and background information questionnaire. Project information

sheets and consent forms are tailored to each project, but share a common format which includes a set of graduated options for consent. Participants can choose whether or not to agree to the collection of different types of data (e.g. recordings, consultation records); whether the data may be permanently archived in the ARCH Corpus or must only be used for the current project; what data formats may be used for research presentation or teaching purposes (e.g. video, audio, or transcript only); and whether they are willing to be contacted again. For certain projects, consent and copyright clearance are also sought to place carefully selected short data clips onto open access website pages.

In terms of data processing, audio-visual recordings are uploaded and catalogued as soon as possible post-recording, and linked to the demographic information entered onto the database and any documents on file, such as the relevant field notes and scanned de-identified medical notes. A research nurse then creates a lay synopsis and timed content log of the video file (including explanation/ glossing of clinical content and terminology) for each recorded interaction prior to transcription by trained research assistants. Our stored base transcriptions use an adapted version of the Wellington Corpus of Spoken New Zealand English conventions, which can be readily converted to Jeffersonian (CA) transcripts, or to text files for automated searching. An interaction summary, including the synopsis, pseudonyms assigned, and key demographic and technical information, can also be generated from the database as required.

The initial investment of time in perfecting a cataloguing system, along with efficient data-processing workflows, manuals and templates to ensure consistency, has definitely paid off, and this data management system continues to stand the test of time. However, it remains quite a manual system, and we are now working with our university research librarians to design a new platform with a view to eventually migrating the archival material and metadata in the ARCH Corpus to a secure online data repository.

3.5 Governance

As the ARCH Corpus has grown and developed into an established research archive, researchers and educators from outside the core

ARCH Group are increasingly seeking to use this body of data for a range of purposes. A Governance Group oversees the development, management, access to and uses of the ARCH Corpus. This currently comprises the four co-directors of the ARCH Group (Maria Stubbe, Tony Dowell, Kevin Dew, and Lindsay Macdonald). Sign off from the ARCH Governance Group is required prior to any access to ARCH Corpus data by researchers not involved in its collection, and for any publications or other outputs drawing on that data. Formal ethical approval from an accredited health research ethics committee is also required for any new research projects. Decisions are made on a consensus basis, with involvement by representatives of the originating project team, where applicable, for an agreed period of time post-completion of that project. The aim is not to unreasonably limit access to the data, but to achieve a balance between meeting the ethical imperatives of maintaining confidentiality/anonymity, data security, respect for contributors and academic fairness, and making the best possible use of this valuable resource.

In the interests of managing potential risks to privacy, confidentiality and reputation of our research participants and of the ARCH Group and university, the Governance Group tends to take a conservative approach to data security and access in general, and to the video data in particular. Participants have consented to the permanent archiving of their recordings on the basis that the ARCH Corpus is not an open access dataset. The precise terms of consent vary, and the data collected can be of a highly personal nature. In addition, video data of real interactions is not readily de-identifiable (unlike interviews or transcripts, which are routinely anonymised), and with advances in information technology now making it easy to move and copy large digital files, stringent processes are required to ensure confidentiality and prevent accidental 'escape' of the data.

Access to Corpus data by people outside the core ARCH Group research team is restricted to the data subset actually required to complete their project. Where data subsets were collected specifically for a particular project, access is automatically granted to members of the research team involved, and where such data is incorporated into the ARCH Corpus, it will not generally be

available for use by other researchers as part of the wider Corpus until after an agreed period. Everyone who does access the data (researchers and project staff, associate and student researchers) is required to sign and abide by a confidentiality agreement, and must agree to strictly follow agreed data-handling protocols to ensure the confidentiality and security of the data at all times. Our current practice is that even in settings such as one-off professional or academic seminars and workshops we generally play only selected anonymised excerpts, unless we have asked for and received explicit permission to do otherwise from the participants, and we do not circulate full transcripts or other documentary material in such contexts.

The ARCH Governance Group reserves the right to review and exercise sign off on any publications, presentations or other outputs based on data held in the ARCH Corpus to ensure that the conditions of use have been met and that the work is consistent with the overall philosophy of 'appreciative inquiry' and respect for our research participants. Publications and authorship are agreed on a case-by-case basis with all stakeholders.

4. Concluding remarks

The development of any kind of corpus or research archive is always a long-term commitment, and is not an undertaking for the fainthearted. However, it is also immensely rewarding, especially in the context of the 'throwaway culture' that has been part of the health research data landscape for too long. There are definite signs of change in the air. Ethics committees in New Zealand no longer routinely remind researchers that they should plan to destroy personal health data collected as part of research projects after ten years (the minimum period it must be kept). Rather, there is growing acceptance that where the appropriate consents for 'unspecified future use' are in place, this should not happen by default. It is also exciting to see projects like the One in a Million Archive in the UK receive funding specifically to collect data for an institutional corpus of consultation data, which demonstrates the greater value now placed internationally on re-using (expensively acquired) data for secondary analysis.

The ARCH Group has largely achieved what we set out to do 13 years ago. We have shown it is possible to gain consent to collect and permanently archive recordings of sensitive health interactions and related data, to do so in an ethically sound way, and to create a serviceable but inexpensive data management system to keep track of it all. The ARCH Corpus continues to be used for a wide range of applied and theoretical research, including by postgraduate students and research associates in New Zealand and internationally, and we are increasingly using our existing research data and findings as a basis for producing educational resources. As noted earlier, our next challenge is to ensure the technical sustainability of this collection and its governance into the future. We would also like to continue adding data of a sufficient technical standard to be used in multimodal analysis, which is becoming a significant methodology in health interaction research.

More generally, it will be very interesting to watch the development of new health-related corpora and supporting technologies. There will undoubtedly be challenges ahead in terms of research data ethics and governance as governmental and social expectations continue to shift about what kinds of data relating to individuals should be collected and stored and how it may be used. At the same time, there will be fascinating new research questions to explore as the 'internet of things', 'telemedicine' and 'electronic assistants' start to change the face of health encounters.

Notes

1. The development of the ARCH Corpus of Health Interactions has been a truly collaborative effort. First and foremost, this venture would not have been possible without the generosity and trust of the hundreds of individuals who have so generously contributed their recordings and information to be archived for ongoing use in research and education. I would like to particularly acknowledge the practical commitment of the ARCH co-directors, Tony Dowell, Kevin Dew and Lindsay Macdonald, who have worked with me on this project from the beginning, and the invaluable contributions of successive data managers (George Major, Sarah White, Rachel Tester) and Stella Ramage who programmed the customized ACCESS database. Many other colleagues, research assistants, and IT and administrative staff have also contributed over the years. Thanks must also go to the funding bodies who have supported the various component projects, including the Health Research Council of New Zealand, the Royal New Zealand Marsden Fund, Lottery Health Research, Ako Aotearoa, the Royal New Zealand College of General Practitioners and the University of Otago.

References

Applied Research on Communication in Health Group. 2017. ARCH Corpus of Health Interactions. University of Otago. www.otago.ac.nz/wellington/research/arch/corpus/.

Barnes, Rebecca K. 2016. Creating a data archive of GP consultations – the motivations and challenges. A blog by researchers from the Centre for Academic Primary Care. University of Bristol. https://capcbristol.blogs.ilrt.org/2016/04/15/one-in-a-million/.

Barone, Susan. 2012. Seeking narrative coherence: Doctors' elicitations and patients' narratives in medical encounters. Unpublished PhD dissertation. Wellington: Victoria University of Wellington.

Byrne, Patrick S. and B. E. Long. 1976. *Doctors Talking to Patients: A Study of the Verbal Behaviour of General Practitioners Consulting in their Surgeries*. London: Her Majesty's Stationery Office.

Carroll, Katherine, Rick Iedema and Ross Kerridge. 2008. Reshaping ICU ward round practices using video-reflexive ethnography. *Qualitative Health Research* 18,3: 380–390.

Crawford, Paul, Brian Brown and Kevin Harvey. 2014. Corpus linguistics and evidence-based health communication. In Heidi Hamilton and Wen-ying Sylvia Chou (eds) *The Routledge Handbook of Language and Health Communication*. Abingdon, UK: Taylor and Francis. 75–90.

Dick, Bob. 2004. Action research literature: Themes and trends. *Action Research* 2,4: 425–444.

Dowell, Anthony, Maria Stubbe, Kathy Scott-Dowell, Lindsay Macdonald and Kevin Dew. 2013. Talking with the alien: interaction with computers in the GP consultation. *Australian Journal of Primary Health* 29: 275–282. doi: http://dx.doi.org/10.1071/PY13036.

Heritage, John and Douglas W. Maynard. 2006. *Communication in Medical Care: Interaction Between Primary Care Physicians and Patients*. Cambridge: Cambridge University Press.

Hudak, Pamela L., Shannon J. Clark and Geoffrey Raymond. 2012. The omni-relevance of surgery: How medical specialization shapes orthopedic surgeons' treatment recommendations. *Health Communication*. 1–13. doi: 10.1080/10410236.2012.702642.

Jepson, Marcus, Chris Salisbury, Matthew J. Ridd, Chris Metcalfe, Ludivine Garside and Rebecca K. Barnes. 2017. The 'One in a Million' study: Creating a database of UK primary care consultations. *British Journal of General Practice* 67,658: e345–e351.

Mondada, Lorenza. 2016. Operating together: The collective achievement of surgical action. In Sarah J. White and John A. Cartmill (eds) *Communication in Surgical Practice*. London, UK: Equinox. 206–233.

National Ethics Advisory Committee. 2012. *Ethical Guidelines for Observational Studies: Observational Research, Audits and Related Activities*. Revised edition. Wellington: Ministry of Health.

Parry, Ruth, Marco Pino, Christina Faull and Luke Feathers. 2016. Acceptability and design of video-based research on healthcare communication: Evidence and recommendations. *Patient Education and Counseling* 99,8: 1271–1284.

Pope, Charlene and Boyd H. Davis. 2011. Finding a balance: The Carolinas conversation collection. *Corpus Linguistics and Linguistic Theory* 7,1: 143–161.

Robinson, Jeffrey D. 2012. Overall structural organisation. In Jack Sidnell and Tanya Stivers (eds) *The Handbook of Conversation Analysis*. Chichester, UK: John Wiley & Sons. 257–280.

Roter, Debra L. and Susan Larson. 2001. The relationship between residents' and attending physicians' communication during primary care visits: An illustrative use of the Roter Interaction Analysis System. *Health Communication.* 13,1: 33–48.

Royal, Te Ahukaramü Charles. 2017. *Papatüänuku – the Land – Whakapapa and kaupapa.* Te Ara – The Encyclopedia of New Zealand. http://www.TeAra.govt.nz/en/papatuanuku-the-land/page-8.

Sarangi, Srikant. 2004. Towards a communicative mentality in medical and healthcare practice. *Communication & Medicine* 1,1: 1–11.

Swinglehurst, Deborah, Celia Roberts, Shuangyu Li, Orest Weber and Pascal Singy. 2014. Beyond the 'dyad': A qualitative re-evaluation of the changing clinical consultation. *BMJ Open* 4,9. doi: 10.1136/bmjopen-2014-006017.

Wyer, Mary, Rick A. M. Iedema, Su-yin Hor and Lyn Gilbert. 2017. Patient involvement can affect clinicians' perspectives and practices of infection prevention and control: A 'post-qualitative' study using video-reflexive ethnography. *The International Journal of Qualitative Methods* 16,1: 1–10. doi: 10.1177/1609406917690171.

Ziebland, Sue and Andrew Herxheimer. 2008. How patients' experiences contribute to decision making: Illustrations from DIPEx (personal experiences of health and illness). *Journal of Nursing Management* 16,4: 433–439. doi: 10.1111/j.1365-2834.2008.00863.x.