Designing a Multimodal Spoken Component of the Australian National Corpus

Michael Haugh
School of Languages and Linguistics, Griffith University

1. Introduction

Spoken language and interaction lie at the core of human experience. The primary medium of communication is speech, with some estimating the ratio of spoken-written language to be as high as 90%-10% (Cermák, 2009, p. 115).[1] Yet they have remained poor cousins in the building of corpora to date. Not only are spoken corpora much smaller than written corpora (Xiao, 2008), the overwhelming focus in the analysis of spoken corpora has been on textual transcriptions of audio recordings, with the original recordings themselves generally not being widely available (Wichmann, 2008, p. 189). In the most comprehensive, large-scale national corpus to have included a spoken component to date, the British National Corpus, for instance, the ratio of spoken to written language is inversely proportionate to that estimated to reflect actual communicative practice. Moreover, as the original sound recordings are not (widely) available, researchers are limited to analysing textual representations of spoken interaction.

As a result of these constraints, work on spoken corpora has largely focused on the analysis of lexical and grammatical features of spoken language (Wichmann, 2008, p. 189).[2] However, as Adolphs and Carter (2007) have recently argued "while current corpora allow us to explore multimillion word databases, they fail to represent language and communication beyond the word. This is problematic as social interactions are in fact multimodal, combining both verbal and non-verbal elements" (p.133). The increasing recognition that language needs to be studied *in situ* necessitates the building of multimodal corpora that allow such analyses to be undertaken (Allwood, 2008, p. 223). Yet while building large spoken corpora that are (at least partially) multimodal appears to be a way forward in redressing the relative neglect of spoken language in corpora to date, such endeavours are likely to be enormously time-consuming and expensive if the myriad of challenges facing those who wish to build such corpora are not carefully unpacked in the initial stages of design. The aim of this paper is thus to consider some of the main challenges involved in designing an Australian National Corpus (AusNC) that is multimodal.

The paper begins by outlining what constitutes a multimodal corpus, and drawing a distinction between multimodal text corpora and multimodal spoken corpora, the latter of which is the primary focus in this paper. The case for why a multimodal spoken component of the AusNC is to be favoured over traditional approaches to spoken corpora is then outlined. Some of the key challenges that arise in designing a multimodal spoken corpus are next explored. In light of such a complex array of challenges, it is concluded that the principles outlined in Agile Corpus Creation theory (Voorman & Gut, 2008) constitute the most pragmatic way forward in designing and building a multimodal spoken component of the AusNC.

[1] Cermák (2009) does not specify how he reached this estimation further, although it presumably applies to the averaged, typical adult's use of language over the course of a day.
[2] Although see Rühlemann (2007) for an excellent study of interactional and pragmatic phenomena that draws from the British National Corpus, and Adophs (2008) for a similarly themed study drawing from the CANCODE corpus.

 Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus, ed. Michael Haugh et al., 74-86. Somerville, MA: Cascadilla Proceedings Project.

2. Multimodal corpora

A corpus can be generally defined as a large, structured and machine-readable set of language data and annotations that is representative of a particular language (or variety) (see Peters, this volume). In the case of spoken corpora, however, such criteria are not always met. Many of the so-called corpora used in studies of language-in-use are, in actuality, language datasets rather than corpora per se due to their limited size or lack of machine-readability. Representativeness also often remains an open empirical question in the case of spoken corpora due to the difficulties and expense involved in collecting and annotating spoken language data. Nevertheless, in order to constitute a solid foundation for empirical studies of language use, building a spoken corpus that meets these criteria is crucial.

A multimodal corpus can be further defined as a large, structured and machine-readable set of "language and communication-related material drawing on more than one modality" (Allwood, 2008, p. 208). This adds an additional layer of complexity to corpus design and building in that multimodal corpora involve not only text as commonly found in traditional corpora, but other sensory (for example, hearing) and production (for example, gesture and speech) modalities (Allwood, 2008, p. 208). As a consequence of these additional modalities, there are arguably two quite distinct types of multimodal corpora in existence. Multimodal *text* corpora involve collections of written data where non-textual elements are considered to be an inherent part of the corpus. Collections of (web-based) text with associated pictures or diagrams are multimodal (Baldry & Thibault, 2006), as are many computer-mediated forms of communication such as e-mail and (we)blogs (Beiβwenger & Storrer, 2008; King, 2009; Ooi, 2009). Multimodal *spoken* corpora, on the other hand, involve collections of spoken data where paralinguistic and non-verbal elements are considered an inherent part of the corpus. These generally involve searchable collections of audio or audiovisual recordings linked with either transcriptions (more commonly), or annotations (more recently) (Knight, Evans, Carter & Adolphs, 2009). The focus in this paper is on the latter type, although the importance of the former should not be underestimated.

Another important distinction needing to be drawn is that between multimodal spoken corpora and speech databases. A prototypical spoken corpus involves recordings of spontaneous interactions between two or more participants in natural settings (Cermák, 2009, p. 117). A prototypical speech database, on the other hand, involves highly constrained or even scripted language in a dedicated environment free from extraneous noise (Wichmann, 2008, p. 187). These conditions are necessary for producing clean recordings that are amenable to instrumental analysis and for developing technological applications of speech research (Wichmann, 2008, pp. 187-188), but the language data collected under such conditions are not very representative of how people really communicate in natural conditions, and consequently less directly relevant to many branches of linguistics, and humanities more broadly. Yet while (multimodal) spoken corpora and speech databases encompass largely distinct bodies of language data, there is nevertheless some potential overlap between less prototypical speech types found in spoken corpora (particularly monologue and elicited spontaneous speech), and thus a certain degree of intersection between them. The content of a multimodal spoken component of the AusNC therefore needs to be designed to be complementary to, and where possible overlapping with the Australian Speech Science Infrastructure (ASSI) or Big Australian English Corpus (Big ASC) (see Burnham et al, this volume). There is potential for further complementarity between the AusNC and the ASSI in relation to their underlying infrastructure, which could largely be shared.

3. The case for a multimodal spoken component of the AusNC

Spoken corpora have traditionally been composed of textual transcriptions of audio recordings that themselves are not generally widely available. The transcriptions that constitute various spoken corpora differ in their level of detail. Some only include what is said (largely in standard orthography), such as in the spoken component of the Corpus of Contemporary American English (Davies, 2009), while others, such as the spoken component of the British National Corpus (Crowdy, 1994), include details relating to turn-taking and timing (including speaker turns, pauses, hesitations, false starts, overlapping speech, and repetition), a limited amount of paralinguistic information (including markedly louder or softer volume and non-verbal sounds such as laughter or coughs), and restricted contextual information. However, as Mulder, Penry Williams & Loakes (this volume) point out, more detailed transcriptions

and access to the sound recordings themselves is crucial to much of the research that is carried out on spoken language. It is thus argued in this section that the development of a multimodal spoken component of the AusNC, where the original audio or audiovisual recordings accompany transcriptions and annotations is essential for moving the research agenda forward in a number of fields within linguistics and language technology, and indeed the humanities and social sciences more broadly.

A multimodal spoken corpus has a number of advantages over traditional spoken corpora. By closely aligning transcriptions or annotations with the original audio or audiovisual media file, a much richer dataset that has broader applications to various language-related fields is created. Rather than being restricted to the analytical interests of corpus linguists, a multimodal spoken corpus is potentially useful to many other fields in linguistics, including pragmatics, conversation analysis, discourse analysis, sociolinguistics,[3] as well as language technologists working on speech recognition, audio(visual) file search technologies, and in some cases, natural language processing. Such a richly layered dataset also gives the researcher much greater analytical purchase. As Allwood (2008: 211) has recently argued, access to richly annotated prosodic information gives analysts considerable insight into information structure and emotions and attitudes, while providing annotations of nonverbal elements of communication, such as gesture, posture, and facial expressions, gives the researcher insight into both emotions/attitudes and communication management (see also Dahlmann and Adolphs, 2009).

The importance of such prosodic information can be seen in an analysis of instances of jocular or non-serious mockery between Australian speakers of English (Haugh forthcoming). If one compares a basic transcription of an interaction with a more richly transcribed version it becomes clear that the former is not sufficient for the analyst to establish that joculary mockery has indeed arisen in an interaction. In the following excerpt taken from a collection of audio recordings of interactions between Australians who are getting acquainted, a candidate instance of jocular mockery can be located in line 326. The interaction is around 14 minutes into a conversation between a female Australian in her early thirties (Emma) and a male Australian in his late twenties (Chris). In the conversation preceding this excerpt Emma has been talking about the acupuncture *sensei* (teacher) she had when she was living in Japan.

(1) AGA: ERCH: 14:54 [basic transcription]
318 E: The sensei that I ended up
319 C: oh
320 E: studying with, the older one
321 C: Yeah
322 E: has been practicing for thirty five years
323 he's a Japanese man
324 C: Wow
325 E: Mmm at least I think he's the best
326 C: Does he do bonsai as well?
327 E: No not like that.
328 ((laughs)) Nup.

Emma at this point in the conversation uses the term *sensei* to refer to her teacher (line 318), which Chris is aware (from his limited knowledge of Japanese) is a sign of respect, and then goes on to praise her teacher as very experienced (line 322) and the "the best" (line 325). However, while Chris at first indicates he is impressed (line 324) in response to Emma emphasising that her teacher is "the best" (lines 325), [4] he goes on to ask whether Emma's teacher can do *bonsai* (that is, grow miniature trees) (line 326), thereby implying that it seems from Emma's account that her teacher can do anything. By associating expertise in acupuncture with expertise in something totally unrelated, Chris frames her extolling as "overdoing" (Drew, 1987), and so mocks her for taking her admiration for the teacher too

[3] And to some extent phonetics and phonology, although the requirement in these fields for very high quality recordings that allow instrumental analysis means such researchers would need to draw from the ASSI as well.
[4] This interpretation of "wow" as displaying that Chris is impressed draws from his intonation in uttering it, as well as Emma's subsequent response in line 325, with which an interpretation of "wow" as sarcastic would not be coherent.

far, or too seriously (what is colloquially known as *taking the piss*). However, from this basic transcription it is not entirely clear whether this is taken to be non-serious mocking or otherwise by Emma, as while she denies it and laughs (lines 327-238), we cannot tell from this transcript whether this laughter receipts the comment by Chris as jocular, or alternatively indicates her annoyance.

A more detailed transcription of the final part of the excerpt above that includes additional prosodic information, however, provides evidence to the analyst that what we have here is indeed an instance of jocular mockery.[5]

(2) AGA: ERCH: 14:54 [more detailed transcription]
325 E: Mmm >at least< I think he's the *best*
(0.5)
326 C: Does he do bonsai as well?
(0.6)
327 E: NO::: NOT LIKE ↓THA(H)T.
(1.2)
328 E: A ha ha ha ha ha .hhh nup

In analysing more closely the way in which Chris delivers his question about *bonsai* and Emma's response it can be established that his question is oriented to by both Chris and Emma as jocular mockery. The question is interpretable as ironic due to its contrast with known facts (that is, acupuncture and growing *bonsai* are completely unrelated fields). However, it is through its delivery with a flat, deadpan intonation by Chris (line 326), as well as Emma's receipting of it with an exaggerated (higher volume) rejection infused with laughter (line 327) and subsequent laughter (line 328), that it is interactionally achieved as jocular or non-serious mockery (Haugh, forthcoming). Without access to these prosodic details, the analyst is left guessing as to whether the interactants are orienting to the mockery as serious or non-serious.

The way in which nonverbal elements can yield a richer, more empirically grounded analysis of another interactional practice, jocular abuse (Haugh and Bousfield, forthcoming), is evident in the following excerpt from an audiovisual recording, where one Australian male in his early twenties has called his sharehouse mate, who is also an Australian male in his early twenties, a "nobhead" (line 28).

(3) GCSAusE06: 1:03
23 N: so you were born
24 on Sunday, (0.5) of the fir:st month, (0.5) of (.)
25 the twenty-seventh day of nineteen eighty three=
26 D: =↑no:, not ↑February ma:n
27 (0.2)
28 N: oh, yo:u're a nobhea:d.
29 (0.6)
30 D: °what° (.) h ha ↑hehehehe .hhhh

Up to this point in the conversation, Nick has been showing David how his new mobile phone can be used to calculate the day of the week on which David was born, which turns out to be a Sunday (lines 23-25). The insult in line 28 is occasioned by David's slipup in thinking the first month of the year is February (line 26). David responds after a brief pause by asking "what" (orienting to Nick's insult in line 28), before displaying realization through his laughter that he has made a mistake. However, while David receipts this abuse as non-serious, it is not entirely clear from Nick's intonation in line 28 whether he himself is orienting to his insult as jocular. In order to establish that *jocular* abuse has been interactionally achieved here, that is, both interactants are orienting to the insult as non-serious, careful examination of Nick's facial expressions and head movements at this point in the interaction prove to be analytically useful. In particular, while David is laughing (line 30), Nick can be observed to

[5] A list of the transcription symbols used in this and subsequent examples can be found at the end of this paper.

momentarily shift his attention from calculating David's day of birth on his mobile by fractionally tilting his head backwards and giving a slight smile (compare figures 1 and 2 below).[6]

Figure 1: 1:16:15

Figure 2: 1:17:03

In doing so, Nick arguably nonverbally frames (and receipts) the insult as non-serious, and thus "nobhead" in this case is interpretable as an instance of jocular abuse. In other words, paying careful attention to nonverbal elements of communication, while not necessarily providing definitive evidence for analytical claims, can nevertheless yield a richer, more nuanced analysis.

Another advantage of creating a large collection of spoken interactions is that it would allow researchers to move beyond demonstrating the recurrence of particular patterns or actions to establishing their generalizability. As Archer, Culpeper and Davies (2008) have recently argued, most studies in pragmatics, for instance, involve "relatively small-scale qualitative analyses that concentrate on small samples of spoken language data" (p.614). While such studies may adequately demonstrate that a particular speech or interactional practice, such as jocular mockery (Haugh, forthcoming), is recurrent amongst a group of speakers, to make generalizations across a particular population requires recourse to much larger tracts of data. A multimodal spoken component of the AusNC would allow researchers investigating languages in Australia to do just that. In this way, we would also "enable analysis of the contemporary nature of language and the identification of empirical patterns that contrast with and frequently contradict traditional descriptions of the organised nature of language" (Adolphs and Carter 2007: 134). While close examinations of spoken interaction in various fields of linguistics over the past forty years have clearly demonstrated that what people ordinarily think about interaction is not always borne out in reality, there is much that remains to be done to further our understanding of spoken interaction.

A multimodal spoken component of the AusNC would also provide an exemplar for a stronger empirical grounding of language use sciences more generally. In many published studies of spoken interaction thus far there has been markedly unequal access to data between those who have undertaken the research, and those interested in the results of that research. While the researcher has full access to the original recordings, those reading published research are generally reliant on transcripts of those recordings provided by the researcher.[7] Problems arise from this practice when the transcripts provided are not necessarily sufficiently detailed, as the grounding of the analyst's interpretations in the details of interaction is dependant on the quality of the transcriptions or annotations utilised, which, moreover, can themselves be disputed depending upon the researcher's own theoretical and methodological stance. It is a generally accepted principle in many fields of scientific enquiry that data used to support claims should be either replicable, or relevant data should be available for interrogation. Making the original media files (whether audio or audiovisual recordings) available for inspection by other researchers, particularly in the process of academic reviewing of work for publication, is one way in which such issues can be dealt with more empirically. However, without adequate research infrastructure, it is likely to be difficult in reality for individual researchers to meet this ideal. A multimodal spoken component of the AusNC could facilitate such a process, as researchers would be

[6] Allowing access to view the original audiovisual recording would, of course, be preferable in this instance. See below for further discussion of this issue.

[7] One exception to this practice can be found in the work of some conversation analysts, where web-based access to original recordings is provided. Emanuel Schegloff's website constitutes an exemplar of this practice (http://www.sscnet.ucla.edu/soc/faculty/schegloff/sound-clips.html), although the widespread adoption of this practice is obviously constrained by the resources of individual researchers.

able to cite original media files and associated transcriptions or annotations in their publications. A multimodal AusNC would thus provide the infrastructure necessary for a more empirically grounded science of language use.

The builders of the largest and most comprehensive spoken corpus to date, the spoken component of the British National Corpus, were constrained by the technologies available to them in the late 1980s and early 1990s. The movement into the hands of ordinary researchers of powerful technologies for digitizing and managing audio(visual) recordings, as well as transcribing or annotating such recordings, heralds a new age for the ways in which we study spoken interaction. The development of next generation multimodal corpora which capitalize on these technological changes, such as the Nottingham Multimodal Corpus (Knight, Evans, Carter & Adolphs, 2009) or the METU Spoken Turkish Corpus (Karadaş & Ruhi, 2009), are arguably part of a broader shift towards greater application of ICT (information and communication technologies) to scientific research, or what is more commonly known as eResearch (Treloar and Wilkinson, 2008). An AusNC that is grounded in principles of eResearch would enable greater sharing of multimodal spoken language data, which is generally very time-consuming to collect, thereby facilitating greater collaboration and interdisciplinarity in our analyses of spoken interaction. At the same time it could encourage more dynamic, two-way flows of data between users of the corpus and the corpus itself, rather than the traditional one-way flow of data. Yet while the application of such technologies to analyses of spoken language data show considerable promise, a number of complex issues arise in considering how we might successfully design a dynamic, multimodal spoken component of the AusNC. A number of these key issues are discussed in the following section.

4. Issues in designing a multimodal spoken component of the AusNC

In this section, it is suggested that in designing the multimodal spoken component of the AusNC three broad issues need to be considered. First, the various ways in which multimodal spoken language data can be represented in the corpus are discussed (see also Mulder, Penry Williams & Loakes, this volume). The need for more sophisticated ways of searching and analysing such data are then considered. Finally, the potential sources of multimodal spoken language data from which we might draw upon are briefly outlined, with a particular focus on how we might make decisions about what (or what not) to include in the AusNC.

4.1. Representing multimodal spoken data

In considering how multimodal spoken data might best be made available in the AusNC, it is useful to draw a distinction between media files (i.e., the original audio or audiovisual recordings), annotations and transcriptions (i.e., textual representations or descriptions of linked media files), and metadata (i.e., information about the creation of the media files and annotations/transcriptions, as well as the participants) (Cassidy, 2008). While each of these levels of data representation encompass particular issues, in all three cases the aim should be to generate machine-readable and standardised data files, thereby building an AusNC that has the potential for interoperability both with other equivalent corpora held in Australia and overseas, and with future renditions of the AusNC itself.

In the case of media files, a number of largely interchangeable standard formats for digitised audio and audiovisual files have emerged. Yet while various standards exist, a key issue facing the designers of the spoken component of the AusNC is that there are both compressed (for example, mp3) and uncompressed (for example, wav) media file formats. While converting uncompressed to compressed file formats is relatively straightforward, the reverse is generally not possible, since vital data is generally already lost in compression processes (Wittenburg, 2008). The choice between compressed and uncompressed formats is thus not a trivial decision, as compressed file formats are considered problematic for instrumental prosodic or phonetic analyses (Deppermann & Schütte, 2008, p. 183; Wittenburg, 2008, pp. 669-672). However, since it is likely that many recordings of interactions made so far have been collected in compressed file formats (although where media files are already in uncompressed formats they should obviously be kept in their original uncompressed formats), it appears that restricting media files in the AusNC to uncompressed file formats is not a realistic aim. Prescribing the use of uncompressed file formats for future recordings, on the other hand, is a distinct

possibility.[8] As a general rule, then, uncompressed media file formats could be considered preferable, but not mandatory, particularly in the case of legacy recordings, for inclusion in the AusNC.

Transcriptions and annotations are generally made of audio(visual) recordings in order to allow for the search and analysis of particular linguistic phenomena in those recordings. Transcriptions, which have been commonly used in traditional spoken corpora, involve the representation of speech in textual form, including what is said, syntactic and lexical units, and (some) prosodic features. Annotations, on the other hand, involve text-based pointers, either embedded in the transcriptions or directly linked the media files themselves. For this reason, while transcriptions are often treated as data in their own right by some researchers, the analysis of annotations tends to remain more closely tied to the original recordings. There are three modes of presentation of transcriptions or annotations: vertical, partiture, and column modes. The vertical mode is generally employed by those who use text-processing software like Word to create transcriptions (with or without annotations), while the partiture mode (which is similar to an orchestral score) and the column mode are utilised in annotation editors, although the latter can generally be converted into the vertical mode (and thus essentially transcriptions) when required. The problem with the still fairly widespread practice of creating vertical transcriptions in programs like Word is that such documents are not readily machine-readable, nor are they easily time-aligned with the original recordings. The use of more specialised annotation editors such as EXMARaLDA or ELAN allows the analyst to create machine-readable annotations with varying degrees of interoperability across different systems, particularly those that store annotations in XML files (Rohlfing et al, 2006), a scenario that is much more amenable to the creation (and use) of a multimodal spoken AusNC.

However, while the increasing interoperability and user friendliness of various annotation editors makes the choice of software increasingly simply a question of the preferences of individual researchers, no matter which annotation editor is chosen, it does require some investment of time on the part of researchers to gain the basic technical knowledge required to operate them. It is likely for some time to come, then, that some researchers will continue to eschew the use of annotation editors in favour of traditional vertical transcriptions in analysing spoken interaction. However, the ready conversion of annotations into traditional vertical transcriptions means that using annotation editors is not an absolute necessity for those wishing to access a multimodal spoken component of the AusNC. For those who wish to contribute spoken data to the AusNC, however, the use of such annotation editors is clearly desirable. In one sense, then, the building of a multimodal spoken AusNC where annotations are stored primarily as XML files that require support from particular annotation editors constitutes an opportunity to encourage the gradual adoption of ICT-supported approaches to analysing spoken discourse by researchers. Due care will need to be taken, however, to ensure that these annotations are interchangeable, and thus accessible through various different types of annotation editors, since different editors have been developed to meet the specific needs of researchers in different fields. It would clearly be counterproductive if annotation data in a multimodal spoken component of the AusNC was only accessible through one particular kind of software.

Another key difference between transcription and annotation is that while transcription generally conflates what is said with prosodic and contextual information, annotation can be conceptualised as occurring at a number of inter-related levels. The most basic is at the level of what is said. Unfortunately, even at this level of annotation standards can vary widely (Deppermann & Schütte, 2008; Oostdijk & Boves, 2008). Orthographic transcription of what is said, which largely follows standard rules of spelling apart from a small number of non-standard exceptions such as *gonna* and *cos*, has been mostly used in traditional spoken corpora. The reason for this is that allowing literary transcription of what is said, where the spelling of words is altered to reflect the particular accent and emphasis of the speaker in question, generally results in significant divergence in how the same words are transcribed by different people. This subsequently creates problems when attempting to search the corpus for particular words or phrases. The other alternative is phonetic transcription, for example, using IPA, although this approach is generally restricted to speech databases (see Section 2). There is no simple answer as to which transcription system might be most appropriate, although the advantage

[8] It is also worth noting that even audiovisual recordings are still at best only a partial representation of a communicative event (Thompson, 2004, p. 62), and so ethnographic observations made in addition to those recordings are best collected whenever feasible.

of allowing for multiple annotations of the same media file come to the fore here, in that an orthographic transcription might be set as the basic standard for the AusNC, while still accommodating those who wish to judiciously add literary and phonetic transcriptions as annotations where required.

Other levels of annotation include syntactic/lexical tagging (for example, POS tagging), prosodic annotation (Wichmann, 2008), pragmatic annotation (Archer, Culpeper & Davies, 2008), and nonverbal/gestural annotation (Allwood, 2008; Blache, Ferre & Rauzy, 2007). The last two forms of annotation, in particular, are very time-consuming and largely theory dependent.[9] For this reason it not realistic to expect a multimodal spoken component of the AusNC to be fully annotated at all these different levels. At the very least though, audio(visual) files should be annotated for what is said, pauses, turn-taking, and basic parts of speech (POS) (Thompson, 2004, pp. 63-65; cf. Mulder, Penry Williams & Loakes, this volume). If media files were linked to these annotations and made available, then researchers would be able to add additional annotations in line with the goals of their project (Wittenburng, 2008, p. 684), and ideally, those annotations could ultimately be fed back into the AusNC and made available for future users.

In other words, it is proposed here that the AusNC be constructed with a minimal amount of annotation (slim annotation), with additional types of annotations being collaboratively created by users of the AusNC according to their needs. Although allowing annotation to be progressively created in this way diverges from the traditional approach to corpus creation, it is argued that a query-driven approach to corpus building is likely to yield an AusNC that can be put to work quickly, and would also avoid inadvertently building early errors in the annotation process into the whole corpus. In order for this kind of cyclic process model and query-driven approach to succeed (Voorman & Gut, 2008), the employment of stand-off annotation according to established standards is a clear requirement. A number of standards for annotating spoken data now exist, including TEI (Text Encoding and Interchange, http://www.tei-c.org) and XCES (Corpus Encoding Standard for XML, http://www.xces.org/) (Lehmberg & Worner, 2008), although such standards can introduce arbitrariness if not carefully implemented (Thompson, 2004, p. 68). Establishing processes for version control and the auditing of annotations would thus be crucial to ensure that high quality annotations for the AusNC are ultimately created.

In regards to information about the recordings and participants, in contrast, it would be advisable to set out specific requirements for all contributions to the AusNC that follow certain metadata standards (Deppermann & Schütte, 2008, p. 186; see also Musgrave and Cutfield, this volume). A number of related standards now exist, including the Dublin Core (http://dublincore.org), OLAC (Open Language Archive Community, http://www.openarchives.org), and IMDI (ISLE Meta Data Initiative, http://www.mpi.nl/IMDI) (Lehmberg & Worner, 2008), the latter perhaps being the most amenable to spoken corpora. One problem with cross-walking across different metadata schemas is that there can be ambiguities in the way linguistic terms or concepts are used in different linguistic domains. As there remains considerable work to be done in establishing a general linguistic ontology to overcome these issues, it is suggested that, for the meantime, employing one single metadata standard for the whole of the AusNC would at least decrease the frequency of such ambiguities.

4.2. Searching multimodal spoken data

The traditional approach to searching corpora has been to employ string-search types for particular lexical and syntactic forms. These results of these searches are generally presented numerically in the form of frequencies for these particular forms, or as lists of concordances. Such an approach to search, however, is not amenable to many branches of linguistics (Archer, Culpeper & Davies, 2008, pp. 615-616), and particularly in the case of a multimodal spoken corpus. It is evident, then, in designing a next generation corpus we need to go beyond simple word or phrase searches, and allow for much more complex forms of search and automated analysis. For instance, users of a multimodal spoken corpora may wish to undertake concurrent searches of certain intonation patterns with particular words or phrases, or may need to search for non-lexicalised elements (such as laughter, crying, long pauses), or alternatively, may be interested in searching for correlations between particular speech patterns and the

[9] Although all annotation or transcription is ultimately selective and imbued with theory according to Deppermann & Schutte (2008, p. 206).

backgrounds of speakers as it is represented in the metadata. Developing search tools that can cut across media files, annotations, and metadata is thus preferable to tools that can only search each in isolation.

Searching metadata and annotations is largely enabled through clear adherence to standards, although can be compromised by the occurrence of incomplete records. Various kinds of software for searching metadata and annotations are also available, such as the ONZE Miner (Fromont & Hay, 2008), although these tend to be tied to the software in which the annotations were originally created. This means an all-purpose tool that can cross-walk annotations would need to be developed in the case of the AusNC. It is also worth noting that for effective search of annotations (or transcriptions) to be undertaken, they need to be synchronized with the relevant media files (Wichmann, 2008, p. 197). This process of time-alignment is also generally software-dependent, although the increasing interoperability of various programs now available is decreasing the possible constraints of this software-dependence. And as previously noted, literary transcriptions can "create massive problems for researchable databases, because there are no canonical conventions for the literary rendition of dialectical, sociolectal, etc. variants of pronunciation" (Deppermann & Schütte, 2008, p. 208). Thus if they are required by researchers, they should be kept on a separate annotation tier.

Lastly, searching audio(visual) data directly is increasingly becoming a possibility, although such direct searches are still largely constrained by the quality of the original recording (Baker et al, 2009). Transcriptions or annotations of what is said are thus likely to be necessary to support search of multimodal spoken data for some time to come, although bootstrapping semi-automated annotations systems using existing transcriptions (Wichmann, 2008) constitutes a potential third-way for rapidly increasing the amount of spoken data in the AusNC that is searchable.

4.3. Sourcing multimodal spoken data

The question of what (spoken) data should be included in the AusNC is likely to elicit a range of different responses from researchers depending on their particular area of interest. However, since the larger and more diverse the collection of data is (both in terms of the demographics of participants and genres or text types), the more useful the AusNC is likely to be to a larger range of users (Peters, this volume), it is apparent that including as much data as can be gathered is a reasonable strategy to pursue in the first instance, particularly in the case of multimodal spoken data which is generally expensive to collect, and for which only a limited amount of data currently exists for Australian English (let alone other languages in Australia).

One of the key decisions in sourcing multimodal spoken data for the AusNC involves the balance between making recourse to legacy collections as opposed to starting new collections of spoken data. While legacy collections are clearly important both for seeding the AusNC, as well as to enable diachronic studies of language use in Australia, ingesting such collections into the AusNC could prove to be enormously time-consuming as annotations and metadata are not always machine-readable, nor are they necessarily encoded in standard formats (Beal, 2009; Fromont & Hay, 2008; Lehmberg & Worner, 2008). Starting new collections of multimodal spoken data, on the other hand, requires both significant funding and careful planning of best practice for the collection process (Allwood, 2008, p. 215; Deppermann & Schütte, 2008; Thompson, 2004; see also Mulder, Penry Williams & Loakes, this volume). Issues particular to the collection of multimodal spoken corpus data also include the synchronization of audio and visual recordings, and the ever present tension between ensuring high quality recordings, on the one hand, and naturalism on the other (Allwood, 2008, p.216). While the Web is a potential source for multimodal spoken data, it is often limited to either audio(visual) recordings without transcriptions, or transcriptions without associated audio(visual) files. The Contemporary Corpus of American English (Davies, 2009), for instance, has the largest collection of spoken data currently available, but since it consists largely of transcripts of what is said, its potential applications for studies of spoken interaction are somewhat limited, even more so than the British National Corpus where more detailed transcriptions are available (Crowdy, 1994).

Yet whether one decides to draw from particular legacy collections of spoken data, or favours starting new collections, the same three key criteria arguably apply: representativeness, balancedness, and comparability. The first criterion of representativeness encapsulates the extent to which the research using the corpus "can stand proxy for the study of some entire language or variety of a language" (Leech, 2007, p. 135). There are various approaches to how we might achieve

representativeness (Biber, 1993; Leech, 2007), but they generally involve a combination of ensuring a valid demographic sample of participants involved combined with including a range of different text types (or genres) in different contexts. The second criterion, balancedness, involves ensuring the language data included in the AusNC is either proportional to its relative occurrence in an entire language or language variety (Leech, 2007), or encompasses the full range of text types to be found (Biber, 1993). Since we still have surprisingly little information about the extent to which different text types occur in various contexts across different populations in a quantitative sense, the prioritizing of text types in the framework proposed by Cermák (2009) is perhaps useful heuristic for prioritizing the different types of spoken interaction to be included in the AusNC. A third criteria to be considered in deciding what data to include in the AusNC is that of comparability, that is, having the same design as other corpora, varying only in temporal or regional provenance of the language (Leech, 2007, pp.141-142). However, this last criterion presumes that previous spoken corpora have been constructed according to best practice (namely, in line with the first two criteria). It is therefore suggested, consistent with the framework proposed by Cermák (2009), that we seed the multimodal spoken component of the AusNC with more prototypical spoken texts, which are defined as spontaneous interactive dialogue between close and equal interactants in a private and informal settings (Cermák, 2009, p.117), before incorporating more specialised spoken texts.

No matter whether one draws from legacy collections or starts new collections, however, a number of complex legal and ethical issues arise. Two of the key issues that need to be considered in relation to a multimodal spoken component of the AusNC are those of privacy and anonymity. Privacy is primarily a legal issue (Fitzgerald, Pappalardo & Austin, 2008), and involves ensuring that the personal details of participants, such as their full names, addresses, place of work and the like, are not made available to users of the corpus. This can be dealt with relatively straightforwardly by using changing these details in transcriptions, and deleting them from media files. Anonymity, on the other hand, is arguably an ethical issue, the idea being that users of the corpus should not be able to recognise a participant. In the case of audio recordings, and particularly audiovisual recordings, however, this is not a realistic condition. It is simply not possible to ensure the anonymity of participants if audio(visual) recordings are made available. One way in which to deal with this issue is to restrict access to media files to particular researchers, the conditions for access being set by the original custodian of that particular collection of data. In the case of recordings from very small communities such access conditions might be fairly tight. But such restrictions should be balanced with both the greater good of allowing access to a wide range of researchers, and the wishes of participants, who are often quite happy for such recordings to be made widely available. Indeed, it is worth noting that the requirements of ethics committees in many institutions for anonymity (and even destroying data after a certain time period has elapsed) were formulated at a time when data was not generally shared, and have been driven, historically at least, by the very particular requirements of medical research, where ensuring anonymity and data integrity is a perfectly understandable condition. In the case of research into language use across larger populations at least, however, such conditions are arguably not consistent with the wishes of participants themselves in many cases, and constitute an impediment to the sharing of language data and thus ultimately progress in the discipline. It is thus proposed that while the privacy of participants must be protected, that the interests of these participants from an ethical perspective be respected through setting various levels of access to media files. General principles to guide decision-making in relation to setting access levels could be developed based on already existing principles.[10]

Another issue that arises specifically in the case of legacy collections of spoken language data relates to the original conditions under which the data was gathered. While collections where participants knew they were being recorded obviously had the implicit consent at some level from the participants, written informed consent as required for ethical research in recent times was often not sought in the psat. Moreover, the data may have been collected for a particular research project, with no mention of the data being used for future research. From a legal perspective, however, while the privacy of participants must be maintained, as discussed above, the ownership of that collection in

[10] For instance, the Statement of Ethics from the Australian Linguistics Society (1989), available at <http://www.als.asn.au/activities.html#ethics>, or Linguistic Rights of Aboriginal and Islander Communities from *The Conference of the Aboriginal Languages Association* (1984), available at <http://www.latrobe.edu.au/alaa/goodprac.html>.

terms of copyright lies with the researcher who led the collection of that data (Fitzgerald, Pappalardo & Austin, 2008). If participants allowed recordings to be made (as opposed to recordings gathered surreptiously), then, the custodianship of the collection lies with the copyright owner, not the participants themselves. It is therefore ultimately up to the custodian of the collection, not the participants, as to how it might be used in future research, including making that data available to other researchers. From an ethical perspective, one is obligated of course to show respect to those participants. But as discussed above, showing this respect can be largely achieved through differentiated access, particularly to original media files.

The issue of copyright also applies to both legacy and new collections of data. From a legal perspective copyright applies to any *collection* of spoken language data, this copyright residing with the originator of the collection. In order to ensure multimodal spoken data can be freely contributed to and accessed from the AusNC, then, it is important to apply creative commons licensing both when adding collections to the AusNC, and for users of the AusNC. Under this model, copyright of the various collections that comprise the AusNC would reside with the originator of the collection, who would license the use of this collection by others through the AusNC. Such an approach to the legal management of the AusNC is consistent with the distributed model for data in the AusNC proposed by Musgrave and Cutfield (this volume).

5. Conclusion

Australia's spoken language data resources are currently scattered and relatively inaccessible, a situation that contrasts with that of many other nations. There is also little done in the way of coordinating new collections of spoken data so that they occur according to current best practice. While some researchers may be content to collect data for only their own purposes, such an isolationist approach to language data collection is detrimental to the ultimate progress of language use sciences. The AusNC is, in intent, an acknowledgement that language data collected through public funding (i.e., grants from the Australian Research Council or from individual public institutions) is ultimately a public good that should be shared, and thus we have a responsibility as researchers employed by public institutions to ensure that such data is shared, as least as far as possible in light of other ethical and legal constraints.

Care must be taken in designing the AusNC, however, to ensure that the language resources it encompasses enjoy cross disciplinary acceptability and wide circulability (Karadaş & Ruhi, 2009, p. 312). The key to ensuring this is to incorporate widely accepted standards for media file storage, annotation and metadata in the initial design of the AusNC. It has been suggested that due to the complexity and potentially time-consuming nature of building a multimodal spoken component of the AusNC, a cyclical process driven by needs of users, as represented in Agile Corpus Creation theory (Voorman & Gut, 2008), be employed, particularly in regards to making decisions about annotation. While a multimodal spoken corpus has been advocated in this paper, this must be developed within the broader context of the AusNC. The distributed architecture for the AusNC proposed by Cassidy (2008) and Musgrave and Cutfield (this volume) would thus be an ideal environment in which to start building a multimodal spoken corpus.

The AusNC also constitutes an opportunity to encourage the adoption of ICT-supported approaches to collecting, managing and analysing spoken language data. The increasing availability to ordinary researchers of sophisticated technologies for the study of spoken discourse promises a bright future for the field, although in discussing what a comprehensive spoken corpus developed using these latest technologies might look like, Wichmann (2008) concludes that "at present the technology outstrips the linguistics" (p.205). The challenge for the designers of the Australian National Corpus is to reverse this current situation.

Transcriptions symbols (from Jefferson, 2004)

(0.5)	numbers in brackets indicate pause length
(.)	micropause
:	elongation of vowel or consonant sound
-	word cut-off
.	falling or final intonation

?	rising intonation
,	'continuing' intonation
=	latched utterances
underlining	contrastive stress or emphasis
CAPS	markedly louder
° °	markedly soft
.hhh	hearable inbreaths
(h)	hearable aspiration or laughter particles within words
↓ ↑	sharp falling/rising intonation
* *	hearably smiling voice
> <	talk is compressed or rushed

References

Adolphs, Svenja. (2008). *Corpus and context. Investigating pragmatic functions in spoken discourse*. Amsterdam: John Benjamins.

Adolphs, Svenja, & Ronald Carter. (2007). Beyond the word. New challenges in analysing corpora of spoken English. *European Journal of English Studies, 11*(2), 133-146.

Allwood, Jens. (2008). Multimodal corpora. In Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics. An international handbook. Volume 1* (pp. 207-225). Berlin: Mouton de Gruyter.

Archer, Dawn, Jonathan Culpeper, & Matthew Davies. (2008). Pragmatic annotation. In Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics. An international handbook. Volume 1* (pp. 613-642). Berlin: Mouton de Gruyter.

Baker, Janet M., Li Deng, James Glass, Sanjeev Khudanpur, Chin-Hui Lee, Nelson Morgan & Douglas O'Shaughnessy. (2009). Research developments and directions in speech recognition and understanding, Part 1. *IEEE Signal Processing Magazine, 26*(3), 75-80.

Baldry, Anthony, & Paul J. Thibault. (2006). *Multimodal transcription and text analysis* London: Equinox.

Beal, Joan C. (2009). Creating corpora from spoken legacy materials: variation and change meets corpus linguistics. In Antoinette Renouf & Andrew Kehoe (Eds.), *Corpus linguistics: Refinements and reassessments* (Vol. 69, pp. 33-47). Amsterdam: Rodopi.

Beiβwenger, Michael, & Angelika Storrer. (2008). Corpora of computer-mediated communication. In Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics. An international handbook. Volume 1* (pp. 292-309). Berlin: Mouton de Gruyter.

Biber, Douglas. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243-257.

Blache, Phillipe, Gaëlle Ferré & Stéphane Rauzy. (2007). *An XML coding scheme for multimodal corpus annotation.* Paper presented at the Corpus Linguistics Conference (CL2007), University of Birmingham, U.K.

Burnham, Denis et al. (This volume). A blueprint for a comprehensive Australian English auditory-visual speech corpus.

Cassidy, Steve. (2008). Building infrastructure to support collaborative corpus research. Keynote paper presented at the HCSNet Workshop on Designing the Australian National Corpus, UNSW, Sydney, Australia.

Cermák, František. (2009). Spoken corpora design. Their constitutive parameters. *International Journal of Corpus Linguistics, 14*(1), 113-123.

Crowdy, Steve. (1994). Spoken corpus transcription. *Literary and Linguistic Computing, 9*(1), 25-28.

Dahlmann, Irina, & Svenja Adolphs. (2009). Spoken corpus analysis: multimodal approaches to language description. In Paul Baker (Ed.), *Contemporary Corpus Linguistics* (pp. 125-139). London: Continuum.

Davies, Mark. (2009). The 385+ million word *Corpus of Contemporary American English* (1990-2008+). *International Journal of Corpus Linguistics, 14*(2), 159-190.

Deppermann, Arnulf, & Wilfried Schütte. (2008). Data and transcription. In Gerd Antos & Eija Ventola (Eds.), *Handbook of interpersonal communication* (pp. 179-213). Berlin: Mouton de Gruyter.

Drew, Paul. (1987). Po-faced receipts of teases. *Linguistics* 25, 219-253.

Fitzgerald, Anne, Kylie Pappalardo, K & Anthony Austin. (2008). *Practical data management: a legal and policy guide*. Brisbane: Queensland University of Technology.

Fromont, Robert & Jennifer Hay. (2008). ONZE Miner: the development of a browser-based research tool. *Corpora, 3*(2), 173-193.

Haugh, Michael. (Forthcoming). Jocular mockery and face. *Journal of Pragmatics*.

Haugh, Michael, & Derek Bousfield. (In preparation). Jocular face-threatening amongst Australian and British speakers of English.

Karadaş, Derya Çokal, & Şükriye Ruhi. (2009). Features of an internet accessible corpus of spoken Turkish discourse. *Field Surveys, Language Corpora, and Linguistic Infomatics. Working Papers in Linguistics and Language Education* (Vol. 3, pp. 311-320). Tokyo: Tokyo University of Foreign Studies.

King, Brian. (2009). Building and analysing corpora of computer-mediated communication. In Paul Baker (Ed.), *Contemporary corpus linguistics* (pp. 301-320). London: Continuum.

Knight, Dawn, David Evans, Ronald Carter & Svenja Adolphs. (2009). HeadTalk, HandTalk, and the corpus: towards a framework for multi-modal, multi-media corpus development. *Corpora, 4*(1), 1-32.

Leech, Geoffrey. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (Eds.), *Corpus linguistics and the web* (pp. 133-149). Amsterdam: Rodopi.

Lehmberg, Timm, & Kai Worner. (2008). Annotation standards. In Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics. An international handbook. Volume 1* (pp. 484-501). Berlin: Mouton de Gruyter.

Mulder, Jean, Cara Penry Williams, & Deborah Loakes. (This volume). Developing a quality spoken component of the Australian National Corpus.

Musgrave, Simon, & Sarah Cutfield. (This volume). Language documentation and an Australian National Corpus.

Ooi, Vincent B. Y. (2009). Computer-mediated language and corpus linguistics. In Yuji Kawaguchi, Makoto Minegishi & Jacques Durand (Eds.), *Corpus analysis and variation in linguistics* (pp. 103-120). Amsterdam: John Benjamins.

Oostdijk, Nelleke, & Lou Boves. (2008). Preprocessing speech corpora: transcription and phonological annotation. In Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics. An international handbook. Volume 1* (pp. 642-663). Berlin: Mouton de Gruyter.

Peters, Pam. (This volume). The architecture of a multipurpose Australian National Corpus.

Rohlfing, Katharina, Daniel Loehr, Susan Duncan, Amanda Brown, Amy Franklin, Irene Kimbara, Jan-Torsetn Milde, Fey Parrill, Travis Rose, Thomas Schmidt, Han Sloetjes, Alexandra Thies & Sandra Wellinghoff. (2006). Comparison of multimodal annotation tools - workshop report. *Gesprachsforschung - Online-Zeitschrift zur verbalen Interaktion, 7*, 99-123.

Rühlemann, Christoph. (2007). *Conversation in context. A corpus-driven approach*. London: Continuum.

Thompson, Paul. (2004). Spoken language corpora. In Martin Wynne (Ed.), *Developing linguistic corpora: a guide to good practice* (pp. 59-70). Oxford: Oxbow Books.

Treloar, Andrew, & Ross Wilkinson. (2008). Access to data for eResearch: designing the Australian National Data Service. *The International Journal of Digital Curation, 2*(3), 151-158.

Voormann, Holger, & Ulrike Gut. (2008). Agile corpus creation. *Corpus Linguistics and Linguistic Theory, 4*(2), 235-251.

Wichmann, Anne. (2008). Speech corpora and spoken corpora. In Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics. An international handbook. Volume 1* (pp. 187-207). Berlin: Mouton de Gruyter.

Wittenburg, Peter. (2008). Preprocessing multimodal corpora. In Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics. An international handbook. Volume 1* (pp. 664-685). Berlin: Mouton de Gruyter.

Xiao, Richard. (2008). Well-known and influential corpora. In Anke Lüdeling & Merja Kytö (Eds.), *Corpus linguistics. An international handbook. Volume 1* (pp. 383-457). Berlin: Mouton de Gruyter.

Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages

edited by Michael Haugh, Kate Burridge, Jean Mulder, and Pam Peters

Cascadilla Proceedings Project Somerville, MA 2009

ISBN 978-1-57473-435-5 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Haugh, Michael. 2009. Designing a Multimodal Spoken Component of the Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 74-86. Somerville, MA: Cascadilla Proceedings Project.

or:

Haugh, Michael. 2009. Designing a Multimodal Spoken Component of the Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 74-86. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2290.