Unit 5 Case studies

5.4 Corpora: The next generation

Dawn Knight, University of Nottingham

This case study reports on an ESRC-funded research demonstrator project that looks at the planning and development of the third generation of corpus tools: the multi-modal, multi-media corpus. It firstly explores the context for the research, explaining the limitations that current corpora exhibit in the exploration of conversation, beyond the mere spoken word. It continues with a brief exploration of the main phases in the planning, development and analysis of corpora for the future, highlighting requirements and problems encountered at each stage.

Communicating 'beyond the word':
Human communication can be best conceptualised as functioning within a 'highly specialised, evolutionary manifestation of a multimodal gestural complex' (Wilcox 2004: 525). It is composed of a complex network of various direct and indirect 'semiotic channels' (Brown 1986: 409, also known as 'paralinguistic features': Haiman 1998: 23). Whilst verbal channels 'carry the basic content of the message' non-verbal channels carry the 'relationship or command part of the message' (Noller, 1984: 7). These channels (modes) of communication differ widely according to their form, function and context-of-use (see Argyle 1975).

In addition, they can both work simultaneously in talk, interacting with and complementing each other, as well as 'counteracting' each other (Maynard 1987: 590). As a result these channels have the potential to differ dramatically within and across conversation boundaries; such natural variability makes it difficult to effectively explore discourse. Perhaps the most revealing methodological approach that can be utilised to explore and analyse such features of language-in-use, in more detail is found in corpus linguistics (CL hereafter).

However, current CL methods can be seen to be limited in terms of the extent that they can be utilised to explore non-verbal aspects of talk. This is because corpora present all channels of discourse in the same format: that of text. This does not cause problems for written language; however, in order to make spoken corpora 'usable', recordings require transcription (a process which can in itself be seen as a pseudo-analysis of speech to create 'data', in other words *representations* of talk.

Saferstein highlights (2004: 213) 'the reflexivity of gesture, movement and setting is difficult to express in a transcript', consequently, in such a format the 'data' lose many of the characteristics (particularly non-verbal) that existed in the spoken discourse itself, characteristics which can play an integral part in determining the function and meaning of linguistic units in conversation. This means the data is, to a certain extent, far removed from the reality of the spoken word. Thus, such corpora are limited in terms of their use for exploring aspects of language-in-use that exist beyond the text.

Corpora for the future:
In order to develop the scope of what a corpus informs us of, the Digital Record research project (funded by the ESRC) aims to outline the blueprints for a new generation of multi-modal corpora. This is essentially a corpus tool with multimedia facilities that provides the capacity for exploring multiple modes of data (extracted from natural language communication contexts), i.e. streamed verbal, visual and audio data, within the same user-friendly interface. Such a tool will, therefore, facilitate the analysis of relationships between specific search terms and parts of speech (the 'linguistic' features) and proxemic movements and gesture ('paralinguistic' features see Haiman 1998: 23) and how such relationships contribute to or regulate the message of the utterance.

The research project is inter-disciplinary, using the expertise of computer scientists, psychologists and linguists. The project design and development uses an approach based on Greenbaum and Kyng's (1991) notion of 'cooperative prototyping', which focuses upon co-designing technical and computational applications and tools which meet the requirements of, and are of use to, a wider research community. The project is currently in its second phase, that of preliminary development and it is hoped that the corpus tool will reach its final phase in 2008, with the aim of making it publicly available some time after this date.

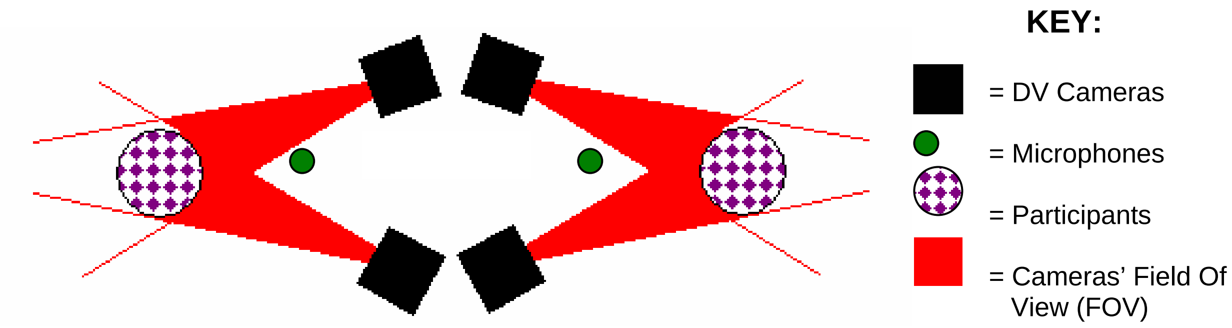
Compiling the corpus:
In order to achieve this, different and more advanced technological procedures for the collection of multiple-modes of data need to be explored in order to develop an integrated way of annotating different aspects of communicative events. To systematically approach the compilation of this corpus a multiple perspective methodology has been used, focusing upon three individual phases:

- **RECORD**: Data sources and collection methods
- **CODE**: Detect, define, encode head nods
- **(RE)PRESENT**: (Re)presentation of data within an interactive, user-friendly interface

Record:
The record stage concerns the sources and collection methods used to compile the data comprising the body of the corpus. Due to the fact that, at present, there has been little corpus research that uses multiple media based tracks of communication, i.e. comprising audio, video and textual renderings of conversation, this project requires the recording of completely new and relevant data sets. Such data need to be rich enough in terms of the information required for in-depth linguistic enquiry, and of a high enough quality to be adapted for successful use with the vision recognition system (discussed in more detail at the code stage). Thus, it was vital as an *a priori* to recording, that various conditions and techniques for data collection were explored and experimented with, to determine those that would achieve the aims set out at the start of the research.

Obviously for such a high-tech corpus, traditional methods for recording, involving the use of tape recorders and Dictaphones alone would not suffice. Instead digital video (DV) cameras were used in conjunction with external microphones. This method provided easy-to-access, high quality audio and visual data, although has a problematic limitation in the fact that such equipment imposes restrictions on how and where recording sessions can take place.

Current vision recognition technologies require us to record each participant face on, with a clear view of head and upper body gestures and movements in order to produce a tractable image for recognition and analysis. Due to this requirement, thus far, data has been collected from dyadic and single speaker contexts only, to allow for the use of (multiple) stationary cameras to collect the best image data possible. So the following set-up has been used:



Participants face each other, with four cameras angled towards them and two microphones situated on the floor between them. This gives four images of the conversation, two of each participant, from a near to face-on perspective, which are displayed in split screens (using a digital multiplexer). The cameras have been positioned to ensure that they provide the highest quality possible in a dyadic situation, which can then be adapted for use with the vision recognition system.

The problem with such a method is that it is arguably not the most 'natural' context for communication. Indeed it is difficult to promote real, 'naturalistic' talk in research settings (although indeed the concept of 'natural' data is itself difficult to define) and in this case, in particular, speakers can feel uneasy about being recorded due to the obtrusive nature of video cameras. However, it is obviously not ethical to 'hide' cameras without consent. In order to minimise this problem all recordings take place in relaxed, familiar settings and since each conversation lasts 45-60 minutes, speakers may become more at ease around recording equipment, promoting talk that is as natural as possible.

In terms of the actual participants recorded for the corpus, it is apparent that intra- and cross-cultural differences, to an arguable extent, can influence the way in which individuals gesture and converse in specific discourse contexts. As a result, for the ease of transferability and consistency, the training data collected for the corpus involved only native English language speakers, in academic environments. Over time, when recognition and coding systems are sufficiently 'trained' and developed, multi-modal data can be gathered from different speakers in different discourse contexts.

Code:
The second stage in the research methodology concerns the methods of data classification and synthesis. Again, the aim was to create an 'intelligent' corpus, a tool that models meaningful gesture-in-talk. Such a tool requires the ability to monitor the function, timing, meaning and response (if any) of the verbal and non-verbal behaviour of all participants, to gain an increased understanding of their significance. Therefore, at the coding stage a feature extractor, i.e. a vision recognition system developed by the computer scientists, is used with classification methodology to extract and label phenomena. These should (adapted from El Kaliouby 2004: 2):

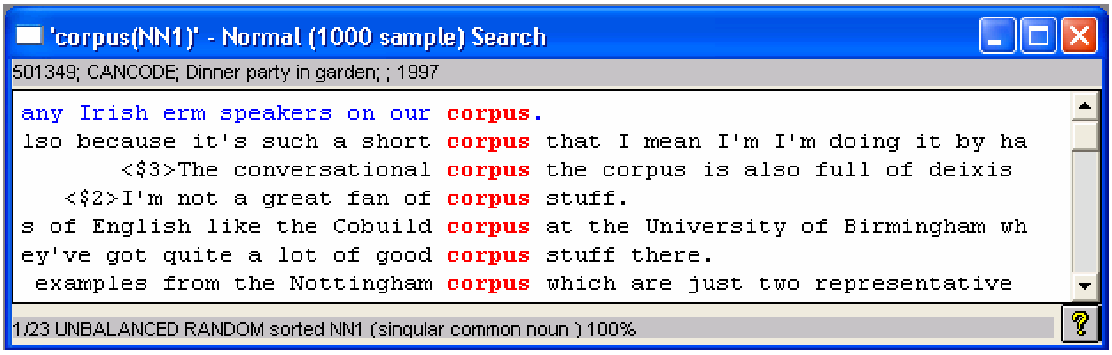
- ✓ **Be dynamic –** with multi-media capabilities
- ✓ **Deal with multiple interacting processes –** i.e. multiple-modes of communication data in synchronicity and with accuracy
- ✓ **Be able to model multi-level abstractions –** i.e. conversation in dyads or groups of participants

There was an intention to make this stage fully automatic but ,as Kapoor & Picard point out in their development of a 'real-time detection' and classification tool, for even salient gestures such as head nods and shakes, this is difficult (2001) due to fact gestures vary widely. It is unlikely, for example, that two head nods are identical, instead they differ according to who is nodding, the intensity, meaning and even how the head is rotated, whether it is simply a rigid up and down movement or whether there is more of an up and slight 20 degree rotation and so on. For Kapoor & Picard (2001: 1) this was problematic because such inconsistent movements caused 'many video based trackers to fail'.

This problem is relevant to all meaningful non-verbal aspects of talk. Obviously, it would not be viable to attempt to control the range of movements used by each participant, to create easy-to-define, discrete units, especially if 'natural' talk is encouraged. Instead it is viable to attempt to account for the range of motions, a challenge that current computer vision techniques find difficult to overcome. Therefore, it was decided that using semi-automatic detection and coding techniques was more appropriate, utilizing the expertise of the linguists with the technical and image know-how of the computer scientists.

In order to create such a coding scheme, it was appropriate to carry out some preliminary linguistic analysis and classification of each stream of data, and then compare findings to determine patterns that may occur both within and across each data stream. The findings are compared with the scientific image analyses to define basic parameters for gesture-in-talk for use as a corpus coding scheme. This iterative process will continue to be used throughout the corpus development.

(Re)present:
The final concern of the corpus development is how the multiple streams of coded data are physically re-presented in a re-usable operational interface format. Current corpora present data in a concordance list format, as follows:


'corpus(NN1)' - Normal (1000 sample) Search
501349; CANCODE; Dinner party in garden; ; 1997
1/23 UNBALANCED RANDOM sorted NN1 (singular common noun) 100%

```
any Irish erm speakers on our corpus.
lso because it's such a short corpus that I mean I'm I'm doing it by ha
       <$3>The conversational corpus the corpus is also full of deixis
   <$2>I'm not a great fan of corpus stuff.
s of English like the Cobuild corpus at the University of Birmingham wh
ey've got quite a lot of good corpus stuff there.
 examples from the Nottingham corpus which are just two representative
```

At the click of a button, appropriate citations of speaker information, context of use and evidence of the specific conversation in which each instance occurs, are easily available. With a multi-media corpus it would be more difficult to exhibit all features of the talk simultaneously, as, if all characteristics of specific instances where a word, phrase or coded gestures (in the video) occur in talk are displayed, the corpus would involve multiple windows of data with, for example 1000 instances of a head nod with an associated audio track of a *mmm* verbalised backchannel and the textual rendering of such. This would make the corpus confusing and impractical.

The basic solution to this problem is to present the data, as with current corpora, in a 'textured' way, with windows of specific and integrate relevant information, layering it behind main frames that display the key search features in a similar way to current textual concordances. With searches of the visual and audio information this is more complex as it is difficult to 'read' multiple tracks of such data simultaneously, as current corpora allow with text. There is an aim to create a balance between the amount of texture, i.e. the complexity and amount of information held in the corpus, and its ease of use. This stage is still in development.

Summary:

There are various technical and practical issues that need to be considered and explored in the development of a corpus tool. This case study has outlined the main phases of development of the next generation of corpus tools, corpora that are based upon multi-media renderings of naturally occurring conversation. This tool will enable the exploration of language-in-use from a more multi-modal perspective, extending the potential for empirical linguistic enquiry. It moves away from traditional corpus linguistic techniques that focus upon the text of talk, giving the analytic tools for exploring the lexical, grammatical and accompanying gestural elements that are part of everyday conversation.

References:

Argyle, M. (1975) *Bodily Communication*. London: Methuen

Brown, R. (1986) *Social Psychology* New York: Free Press

El Kaliouby, R. & Robinson, P. (2004) Real-Time inference of complex mental states from facial expressions and head gestures *Workshop on Real-Time vision for HCI IEEE conference on computer vision and pattern recognition*. Washington

Greenbaum, J. and Kyng, M. (eds.) (1991) *Design At Work: Cooperative Design of Computer Systems* Hillsdale, NJ: Lawrence Erlbaum Associates

Haiman, J. (1998) The metalinguistics of ordinary language *Evolution of Communication 2*, 1: 117-135

Kapoor, A. & Picard, R.W. (2001) A Real-Time head nod and shake detector. ACM International Conference Proceedings Series. 1-5

Maynard, S.K. (1987) International functions of a nonverbal sign head movement in Japanese dyadic casual conversation *Journal of Pragmatics 11*, 589-606

Noller, P. (1984) *Nonverbal communication and marital interaction* Oxford: Pergamon Press

Saferstein, B. (2004) Digital technology and methodological adaptation: Text on video as a resource for analytical reflexivity *Journal of Applied Linguistics 1* (2): 197–223

Wilcox, S. (2004) Language from gesture *Behavioral and Brain Sciences 27*, 4: 525-526

CANCODE is the Cambridge and Nottingham Corpus of Discourse in English, Property of Cambridge University Press.