

Linking the verbal and visual: new directions for corpus linguistics

Ronald Carter and Svenja Adolphs

School of English Studies, University of Nottingham

Abstract

This paper discusses an ongoing research project to investigate the compilation of a small corpus and the development of appropriate software tools that enable a more multi-modal approach to language data. The research draws on recent experience developed in the development of spoken corpora to explore alignments of the verbal and the visual and, as a starting point, does so with particular reference to gestures in communication and the role of head nods in particular. Issues of appropriate data capture and description are discussed alongside questions about the nature of language necessarily raised by language research that goes beyond the textual.

1. Introduction

Advances in the field of corpus linguistics over the past two decades have made it possible to develop computerised multi-million word databases of spoken and written language alongside powerful software tools to analyse this data quantitatively and qualitatively, a development that has contributed to pioneering research in many areas of communication studies and language description. However, while the analysis of large-scale text corpora can provide insights into language patterning and can help establish linguistic profiles of particular social contexts, it is limited to the textual dimension of communication. Communication processes are multi-modal in nature and there is now a distinct need for the development of corpora that enable the user to carry out analyses of both the speech and gestures of the participants in a conversation, and of how the verbal and non-verbal complement one another. In other words, corpus linguistics and discourse analysis might begin to be more closely aligned and descriptions made of rich contexts of language use of the kind advocated and illustrated by Michael Stubbs throughout his career.

1.1 Multi-Modal Communication

Recent work in multi-modal communication has seen advances in both theory and practice. The theoretical starting point for much significant work has been systemic-functional linguistics. Systemic linguistics is a theory that focuses on meaning, choice and probability in language and on the significance of language as a social phenomenon, underlining how particular choices of word, grammar and structure encode different meanings in different contexts of language in use.

Foundational work in multi-modal communication such as Kress and van Leeuwen (1996) has illustrated how choices of image can align with verbal choices and this work has been extended in recent years to embrace the multi-modal analyses of word, image and sound within different language varieties, including cartoons, comics, film, information leaflets, maps, advertisements (including TV advertisements), web pages and classroom textbooks (e.g. Baldry and Thibault, 2004, 2006). The emphasis has been on how choices of one image or camera angle or colour tone can cumulatively encode particular meanings. The almost exclusive focus has been on written text.

A particular challenge for current research is therefore to integrate the computer-enabled power of corpus linguistic methods, the theories and practices of multi-modal linguistic research and, with particular reference to the analysis of spoken discourse, the non-verbal signals of human gestures and bodily communication. In other words, one key aim is to provide computerised analyses of patterns of verbal and non-verbal meaning in ways that allow new understandings of textuality to emerge.

1.2 What is a Gesture?

Human communication functions within a variety of direct and indirect 'semiotic channels' (Brown, 1986: 409) which interact with, complement and 'counteract' each other (Maynard, 1987: 590). The occurrence of such channels is affected by modes of communication that differ widely according to their form, function and context-of-use (see foundational work by Argyle, 1969 and Ekman and Friesen 1969, 1976) and more recent studies by Wilcox 2004 and Gu, 2006). However, most studies have been undertaken within a research paradigm of psychology and in experimental rather than naturalistic conditions.

To date, experimental studies of the multi-modal nature of discourse have in general been designed to answer one or both of the following questions (Kendon, 1994: 177):

- 1 If recipients are offered utterances which include gestures and if they are permitted to see these gestures, do they interpret these utterances differently than when they are not permitted to see them? (examples of such studies include (Dobrogaev, 1929, reported in Kendon, 1980; Rogers, 1978; Riseborough, 1981).
- 2 If recipients are asked to make judgements about the gestures of others in the absence of speech to which they were related, do they make such judgements in a consistent way, and, if they do, do these judgements show that they have some understanding of the utterance of which they were a part?

Studies of gesture and the multi-modal nature of communication have focused upon gaze, (see Griffin, 2004 and Beattie & Shovelton, 1999, 2002) hand movements (see Rimé & Schiaratura, 1991 and Thompson & Massaro, 1986), head movements and other related gestures. In these studies the focus tends to be on language use in experimental conditions and does not embrace spontaneous,

natural conversation. In addition, such studies tend to be more concerned with the gesture in relation to the basic content of talk, and do not explicitly explore the links between specific forms of language and accompanying gestures.

Current gesture detection and recognition systems developed in computer science within a tradition of automated vision recognition (see Nixon and Aguado, 2002; Kapoor and Pickard, 2001)) often focus on precise, intentional gestures. This is particularly true of hand gestures, where applications in sign-language recognition and human-computer interaction mean that specific gestures are made that are designed to be clearly distinguished by the observer. Gestures made in authentic, face-to-face conversation, by contrast, are much fuzzier, their form and meaning open to a greater degree of interpretation – a shake of the head can, for example, indicate disagreement, disbelief, or confusion, creating particular challenges for automated analysis of conversational gesture. Gestures are unlikely to be uniquely identifiable and interpretation will need to take into account other cues, such as the current role of the gesturer (speaker/listener) and the co-text of the conversation (i.e., what occurs before and after a sequence of gesture and talk). Furthermore, intentional gestures arise in a more constrained set of situations than conversational gestures. As a result, image sequences are usually acquired from a small, and known, set of viewpoints. Most intentional gesture recognition systems assume that a lone participant is in clear view, facing the camera from a short distance away. Many also assume the background to be uniform and fixed. Real conversational gestures arise in a wide variety of situations and involve dynamic activities from a variety of viewpoints and distances and include multiple participants, cluttered backgrounds and other moving people and objects.

However, for a corpus of gestures to be developed a record of the image is required and current computer technology provides one of the best available means of capturing such images digitally. The next sections report on a corpus-based project to investigate such a phenomenon with a focus on naturally occurring interactive two-party discourse.

2. **Headtalk: an outline**

HeadTalk is the first step in a project based at the University of Nottingham, involving interdisciplinary research between applied linguists and computer scientists, (in particular experts in vision recognition). The project aims to combine both linguistic expertise and new computational techniques and applications to provide the knowledge, research tools and procedures for exploring the behaviour of some salient gestures in naturalistic conversation. An initial focus on head nods was selected on account of their significance in communication.

The *Headtalk* project team has collected to date (January, 2007) five hours of video data, all based on face-to-face conversational episodes involving native English speaking academics and students based at the University of Nottingham.

The participants were filmed face-on, in close proximity to the cameras in order to create high quality, high resolution images, but were filmed in such a way as to minimise the interference and invasiveness of the recording equipment, to make the participants feel at ease and comfortable in the environment and to allow for (relatively) natural, authentic communication. This data can be properly described as 'multi-modal', as the transcribed recordings provide three different modes of discourse, offering three separate streams of data for analysis: the audio, the visual and the textual.

Utilising computer vision technology

The project utilises research in computer vision technology to allow the research team to detect, recognise and extract descriptions of head nod movements. For the detection and extraction of these movements, a variety of techniques were tested on significant samples of the data. After numerous evaluations, a head tracker was developed which can be placed upon the face of image data. Successions of movements can then be monitored and matched to the basic up-down sequence of a head nod in order to define where the movements occur, with the head tracker tracking movement in the videos. The headtracker allows multiple targets to be tracked in parallel, producing a description of the motion of each and showing intermediate results as they are obtained. (For further description of the tracker used (Cvision) see the Acknowledgements to this paper).

Developing linguistic categories

Head nods are vital for conversational maintenance and management (McClave, 2000) and often function as a form of 'back-channel' (Yngve, 1970), that is, a 'mechanism used for feedback' in discourse (Allwood et al, 1992), involving a strategy which involves a form of 'minimal response', a way for the listener to communicate that they have heard and perhaps understood a speaker's message, while allowing the speaker to continue talking. Although there has been research into and analysis of verbal back-channels, for example minimal responses or 'vocalisations' such as *mmm* and *yeah*, (Gardner (1998, 2002) integrated explorations of the verbal and visual components of head nod behaviour of this nature are limited. Preliminary linguistic analyses and classifications of each stream of data, (i.e. the transcribed text of the talk, as well as the audio and the video) was undertaken to determine patterns that may occur both within and across each data stream. The findings were then compared with the computational image analyses to define basic parameters for this particular gesture.

Coding back-channels

One of the key areas of concern of this project is how the head nods should be encoded. In terms of verbal realisations of back-channels most existing schemes focus upon grouping these in terms of their functions in discourse. This is a useful point of categorisation as every back-channel has a function in discourse, even if it may be unconscious to the interlocutor. Indeed, a wealth of research exists which agrees that 'back-channels have more than one macro function' (O'Keeffe and Adolphs forthcoming) as defined below (see also Schegloff, 1982; Maynard,

1987, 1990). As a guide to the key functions, the framework provided by O'Keeffe and Adolphs, has been adopted in the Headtalk project:

- **Continuers:** Maintaining the flow of discourse (see Schegloff, 1982)
- **Convergence tokens:** Marking agreement and disagreement
- **Engaged response tokens:** High level of engagement, with the participant responding on an affective level to the interlocutor.
- **Information receipt tokens:** Marking points of the conversation where adequate information has been received.

While this basic categorisation can be a useful starting point in analysing verbal realisations of back-channels, the question of how verbal and visual realisations interact within and across such categories has remained largely under-explored. For example, a back-channel such as *yeah* or *right* or *I see* or *mm* can be accompanied by a continuum of possibilities ranging from minimal head gesture to an emphatic nod of significant duration. And duration can also comprise several smaller nods within the same unit and still be linked to the same verbal token. Much depends on how an interlocutor is responding, whether he/she is simply maintaining contact or is signaling something altogether more engaged and involved. It is not just verbal form or duration that are significant but such factors as pitch and intensity govern how the form is interpreted and coded in relation to its verbal counterpart. The relationship is a complex and elusive one and a definitive coding scheme is still very much in development and will be extended beyond this phase of the project.

3. Methods

3.1 Record

For ease of transferability and consistency the data involved only native English language speakers taking part in 45-60 minute PhD supervision sessions. This meant factors such as intra and cross-cultural differences, which can potentially influence the way in which individuals gesture or signal feedback, were minimised.

For the recording of the video participants were required to face each other, with 4 cameras angled towards them and two microphones situated on the floor between them. These images are displayed in a split screen and have been positioned to ensure that they provide the highest quality images possible (see figures 1 and 2).

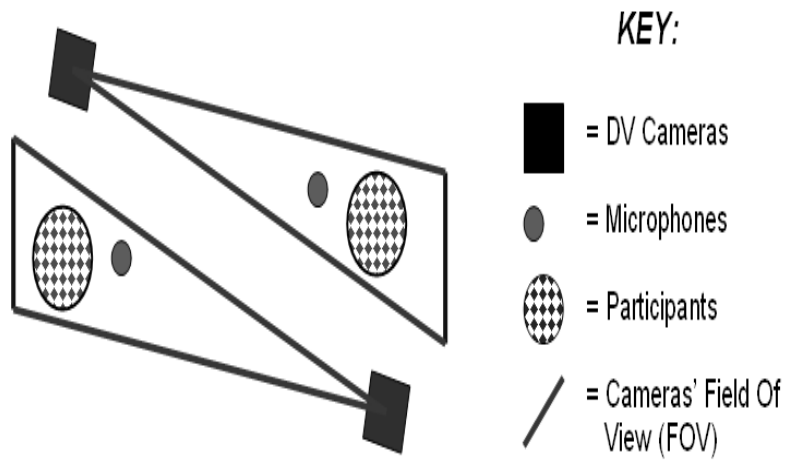


Figure 1: Setup for video recording

In order to keep the images as large as possible, the following screen split for capture was decided upon (figure 2):



Figure 2: Screen shot example of data collected with the modified recording set-up.

Transcription of the data also created challenges as no universal, standardised transcription conventions exist for this type of data. Therefore, for continuity, the conventions used in the CANCODE (Cambridge and Nottingham Corpus of Discourse in English) corpus based at Nottingham University (http://www.cambridge.org/elt/corpus/corpora_cancode.htm) were reapplied here for the purely verbal component with Transana used to allow time stamping and annotation of synchronised video and audio data streams (<http://www.transana.org/>).

3.2 Coding

Following the data collection and rendering of video to display both participants as shown in figure 3, the next step is to develop a coding scheme for verbal and visual signals of active listenership in the data. The main aims of the development of the coding scheme are as follows:

- Defining and classifying verbal back-channel behaviour
- Defining and classifying non-verbal back-channel behaviour
- Combining verbal and non-verbal classifications and highlighting the potential for exploring patterns and relationships between the two

In relation to the coding phase methodological approaches for each of these processes needed to be closely explored. In order to create a new coding scheme, preliminary linguistic and gestural analyses and classifications of the data were undertaken. The findings were cross-compared in order to define basic parameters for gesture-in-talk for use as a corpus coding scheme.

The basic linguistic functions that were used in the analysis of the transcript are those outlined above (O'Keeffe and Adolphs, forthcoming). The analysis of head-nods, on the other hand, is based on classifications established with the use of computer-vision techniques. Five broad types of head-nods were identified in our training data:

Type A: small (low amplitude) nods with short duration

Type B: small (low amplitude), multiple nods with a longer duration than type 1

Type C: intense (high amplitude) nods with a short duration

Type D: intense and multiple nods with a longer duration than type 3

Type E: multiple nods, comprising of a combination of types 1 and 3, with a longer duration than types 1 and 3.

Using the functional categories, as well as the head nod classifications above, we carried out a preliminary analysis of a 10 minute stretch of video extracted from a longer MA supervision session. The participants in the session are a male supervisor and a female student, both of whom are British. The extract was taken from the middle section of the supervision, between 15.00 and 25.00 minutes. The data was transcribed and annotated (see below). The overall word-count of the transcript is 2156 words, of which 1401 words were uttered by the supervisor. For the purpose of illustrating the coding scheme we will focus here only on the description of back-channels used by the supervisor.

The supervisor uses 40 verbal back-channels in total, of which 18 are accompanied by a nod and 22 are purely verbal. In addition the supervisor uses 24 nods which are not accompanied by a verbal signal. Thus, the supervisor uses 42 head nods, 18 of which are accompanied by a verbal signal.

Focus on verbal back-channels

So far, linguistic research has focused mainly on the classification of verbal back-channels as outlined above. When we consider the discourse functions of the 40 verbalised back-channels used by speaker 1, the following breakdown emerges:

Continuer: 11

Convergence Token: 9

Information Receipt Token: 14

Engaged Response: 6

Focus on head-nods

In order to analyse the interface between verbal and visual, we have, as a second step classified the head-nods of the supervisor according to the different criteria (amplitude and duration) that led to the five head-nod types outlined above. Our analysis of the different types of head-nods used by the supervisor generates the following results:

Type A: 13

Type B: 13

Type C: 12

Type D: 2

Type E: 2

Integrating back-channel function and head-nods

We are particularly interested in this analysis to see whether any patterns emerge in those instances where a verbal back-channel is accompanied by a nod. This is the case in 18 of the back-channels used by the supervisor. In terms of linguistic functions and head-nod types the 18 instances are categorised as shown below:

Continuer: 4

Convergence Token: 4

Information Receipt Token: 3

Engaged Response: 7

And:

Type A: 6

Type B: 4

Type C: 6

Type D: 1

Type E: 1

An analysis of back-channel functions as coded with the use of the linguistic coding scheme in relation to the type of nods that co-occur with the different functions highlights a number of interesting trends. Half of the small nods of short duration (type A) co-occurred with the information receipt function, while half of the small nods of longer duration (type B) co-occurred with the function of

a convergence token. All of the type C nods (i.e. short and intense nods) used by the supervisor are accompanied by a verbal signal that has been classified as carrying either the continuer or convergence function. Overall, it is important to take a discourse level perspective to this kind of analysis, as preliminary inspection of the data suggests that some of the functions of head-nods can be aligned with the place at which they occur, i.e. where they are placed vis-à-vis the main speaker's utterance.

These are preliminary results and more data needs to be analysed to see whether there is any stable relationship between head-nods and linguistic signal. However, this very brief illustration of the different coding schemes highlights the need for an integrated analysis of verbal and visual, as the functions of back-channels are modified through the use of head-movement, and it remains to be established whether this modification is one of degree or of kind. One of the main challenges of multi-modal corpus analysis and representation is that corpus linguistics has traditionally focused on discrete items, such as individual words or grammatical categories. The complexities of gesture and movement, on the other hand, mean that they might not be able to be studied alongside traditional corpus linguistic units of analysis in a straightforward manner. Baldry and Thibault (2006: 181) point out that it is 'critically important [...] that corpus-based approaches to text engage with the level of discourse analysis and discourse-level meaning relations on various scalar levels of textual organisation'. While the integration of scalar levels and discrete categories is likely to cause problems in the development of an integrated framework, it also promises to lead to a much richer description of patterns in social interactions.

3.3 Coding the Data: An emerging replay tool

As we have seen, the primary challenge for developing support for analysis of multi-modal corpora is one of developing tools that provide an *integrated* approach to the representation of data. In general, there is a need to create tools that support the 'marking up' or identification of multi-modal patterns and the subsequent codification of recognizable patterns. Coding schemes for marking up textual records and verbal aspects of talk already proliferate. However, there is a paucity of such schemes for handling non-verbal elements: gestures, facial expressions, gaze, head and body movement, posture etc.

Existing tools do not generally support the extraction of linguistic patterns and thus fail, for example, to enable links between different types of listenership and accompanying head movements to be established. There is a need to develop new tools from the ground up to support linguistic analysis and, as an initial step towards this, and by means of developing concrete requirements for technical support, we have sought to exploit an emerging Digital Replay System (DRS) that has been developed to support ethnographic inquiry (Crabtree et al. 2005; French et al, 2006). The Digital Replay System provides some limited mechanisms of representation and below we consider both their potential and limits as a basis for articulating future requirements.

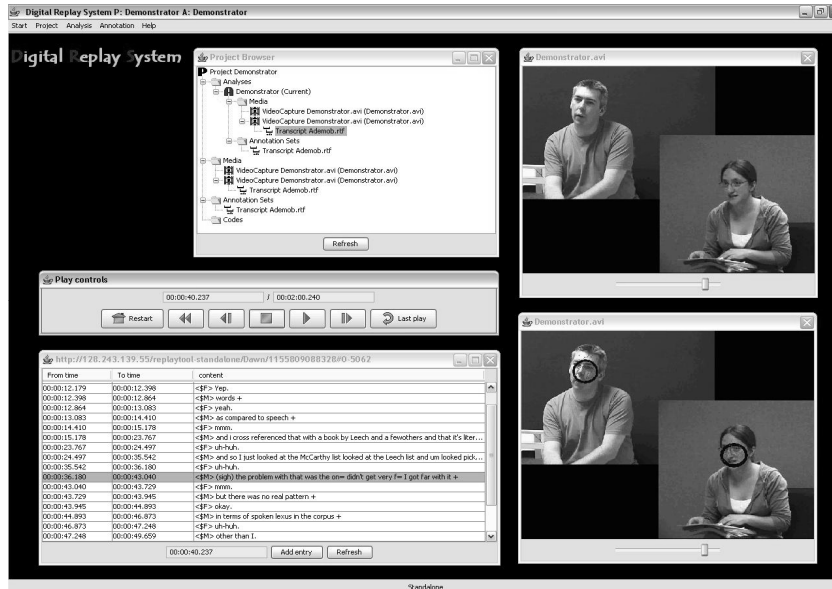


Figure 3: Digital Replay System

The Digital Replay System allows video data to be imported and a digital record to be created that ties sequences of video to a transcribed text log, accompanied, where appropriate, by samples of data that are also tracked by the adopted C-vision recognition system (indicated in blue circles in Figure 3). The text log is linked by time to the video from which the transcript is derived so that the text log plays alongside the video. Further annotations can be added to the log to show where gestures – head nods in this case – occur and these annotations are also tied to the video. An index of annotations is produced and each can be used to go to that part of the log and video at which they occur. The annotation mechanism provides an initial means of marking up multi-modal data and of maintaining the coherence between spoken language and accompanying gestural elements.

The data in the Digital Replay System is presented as a continuous, linear sequence of communication. Yet within any sequence a substantial number of utterances and gestures made by speaker and hearer overlap. This means that the representation of gestural patterns can appear somewhat disjointed, as there is no way at present to represent overlaps. The result of this is that it appears that head nods last only for a specific time and only occur between two verbalisations, which is inaccurate and misleading, as a nod may start after one verbalisation and continue for a long period into the next. Head nods are, in short, variable. They are not fixed in length and, given the limitations of the current incarnation of the Digital Replay System, are difficult to code as events occurring over time and not at particular moments in time. There is, then, a need to develop ways in which their occurrence across utterances can be time stamped and marked out. One way

in which this might be achieved is to represent the utterances and gestural patterns made by parties to a conversation in individual transcripts, so to develop a 'textured text' consisting of separate layers. However, this is not necessarily an ideal solution since head nods need to be represented in relation to the behaviour of both the speaker *and* the listener. Here it is important to identify the defining parameters of the visual aspect of the particular gestural episode and align this with the verbal realisation that may or may not coincide with it.

While there is necessarily a level of interpretation in transcribing spoken discourse, the textual element of the head nod episode is relatively easy to establish and patterns of common back-channels can be extracted from existing multi-million word corpora. These types of patterns have been linked to particular functions in the area of corpus linguistics. For example, minimal verbalisations such as "mhm" have been linked to a continuer function while certain multi-word units such as "that's right" have been linked to an agreement function. Yet, the accompanying head movement, as well as the intonation pattern, can change the function of the back-channel realisation, which in turn will affect the surrounding discourse. It is therefore important to be able to establish some way of recognising the visual elements in a *principled way* so that these can be studied in relation to the verbal elements without adding a burdensome layer of interpretative intervention to the initial representation of the data. There is, thus, a need to marry vision recognition tools to machine learning techniques to reduce the overhead of interpretive work and have these tools and techniques work *across utterances* to adequately represent the verbal-visual character of multi-modal back-channels.

There is thus a need to *marry visual coding schemes to verbal coding schemes*, which may then be exploited by machine learning techniques to codify recognizable multi-modal patterns. In terms of using corpus linguistic techniques to analyse patterns in language it becomes even more important that recognizable patterns are consistently coded with reference to an agreed and replicable coding scheme. If this is not applied throughout the corpus, any searches for patterns will inevitably fail as they rely on the recurrence of consistent representations of linguistic phenomena. Furthermore, coding schemes need to be developed in such a way that they can be shared across different research communities with different community cultures and different representational and analytical needs. In such circumstances, as analytical categories are re-classified in the light of new audio-visual evidence and as new insights from different research communities emerge, coding schemes need to be maintained as dynamically as possible.

3.4 Representation

The final concern of the corpus development is related to the way in which the multiple streams of coded data are physically re-presented. Current corpora utilise concordance tools. 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 is more difficult to exhibit all features of the talk simultaneously. If all characteristics of

the specific instances where a word, phrase or coded gestures (in the video) occur 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 back-channel and the textual rendering of such (as seen in figure 4). This would make the corpus confusing and impractical.

<\$1> Right.	<\$2> Yeah we did have some forms whe	Sound	Video
whatever.	<\\$O1> Yeah. Okay.	Sound	Video
<\$1> Thanks very		Sound	Video
now then?	<\$2> Yeah if you could. If you could pla	Sound	Video
you there?	<\$2> Yeah.	Sound	Video
<\$O2> You are yeah.	<\\$O2> Yeah. Yeah.	<\$2	Sound Video
<\$2> <\$O2> Yeah.	<\\$O2> Mm. Right. Actually	Sound	Video
an hour?	<\$2> Yeah okay.	<\$1> Okay. Bye.	Sound Video

Figure 4: Representing concordances of multi-modal linguistic corpora

The basic solution to this problem is, as with current corpora, to present the data in a 'textured' way and integrate relevant information, such as the codes and further annotations, layering it behind main frames that display the key search features in a similar way to current textual concordances. This would involve marking up the transcript data with relevant information, such as codes for different gestures or indeed with information on the function of each gestures as well as corresponding speaker and time stamps, whilst linking it to other frames of information.

When, for example, the code *+NOD+* is selected in the corpus, the user will have access to the video and accompanying audio. Indeed such features may be relatively straightforward when just marking up single gestures (this has been the basic method used so far in our explorations of the supervision data), but, if one were to mark up additional features, this would become even more complex, especially when 'reading' concordances of multiple sources of data. So with searches of the visual and audio information it is difficult to 'read' multiple tracks of such data simultaneously, as is the case with current corpora and text. Our aim is to create a balance between the amount of texture in the corpus, i.e. the complexity and amount of information held in the corpus, and its ease of use. This is still very much under development.

4. Future Research Priorities

There are a number of lines of research arising from this project that require investigation in the future. These include technical issues such as the development of a recognition system to operate over the tracking data, and issues of scope, such as the analysis of other gestures and the analysis of coupled gestures and linguistic accompanying signals, such as hand movements performed in parallel with head gestures. *Headtalk* has allowed us to gain a better understanding of how we may describe and represent multi-modal language data but has also generated a set of additional pertinent research questions in the process. In addition to those outlined above, these also include theoretical questions of how gesture and language integrate and whether they can be described within a single framework. Major theoretical questions in this connection include consideration of the extent to which gestures may be said to be a language in the sense understood of language as a verbal medium. For example:

- Do gestures have rules and if so, how are the boundaries drawn?
- Do gestures have a syntax, that is, are they syntagmatically and/or paradigmatically organised. Or do they not conform to such structuring?
- If the relationship between language and image can be modally connected, as argued by theorists within a systemic linguistic tradition, and if images can be interpreted according to paradigms of choice, is the same true for gestures and for the relationship between human gestures and language?
- Is a system that is different to a linguistic system and are different underlying theories needed to account for the sheer multiplicity of different gestures?
- Many possible instantiations of head nods have been reported in this paper. What happens when researchers begin to try to explain the many possible meanings of hand gestures and their different cultural manifestations?
- What about 'body language' in the sense of movements encoded interactionally by proxemics?

Another important priority for future research in this area is the development of tools and methods to address ethical issues; for example, to anonymise video data while still being able to extract the salient features that are the focus of the analysis. Pixellating faces or using shadow representations of heads and bodies can blur distinctions between gestures and language forms and, when taken to its logical conclusion, anonymisation should also include replacing voices with voice-overs and with other speakers. Ethical considerations of re-using and sharing contextually-sensitive video data as part of a multi-modal corpus resource need to be addressed further in consultation with end users, informants, researchers and ethics advisors. The issues are especially acute when tools are shared or are developed to be web-enabled.

The *Headtalk* project complements core strands of work to be carried out by the e-Social Science Node at the University of Nottingham (see <http://www.ncess.ac.uk/research/sgp/headtalk/>) As an extension to *HeadTalk*, the Digital

Record project, hosted in the e-Social Science Node (see <http://www.ncess.ac.uk/research/nodes/Digital/Record>) allows for conversational gesture recognition and mark-up of a wider range of different gestures, from hand movements to gaze and facial expressions. This will enable researchers to start to 'complete the picture' of communication, to allow them to think about and explore communication from a variety of different perspectives, something for which *Headtalk* has endeavoured to provide the ground.

5. Conclusion

Natural language is a focus of a diverse range of disciplines and the continued explication of its real world, real time organization is of broad interest. The impetus towards multi-modal corpora recognizes that natural language is an embodied phenomenon and that a deeper understanding of the relationship between talk and bodily actions, particular gestures, is required if we are to develop more coherent understandings of the collaborative organization of communication (see also Saferstein, 2004).

Core requirements towards meeting this goal include the development of machine-based techniques that enable all visual and verbal patterns to be aligned and enable common multi-modal patterns to be recognized. There is also a pressing need to integrate visual and verbal coding schemes and to develop techniques whereby these analytic schemes can be exploited in machine learning environments to codify recognizable multi-modal patterns in large corpora of data. In order to achieve these developments we need to gain a better understanding of the particular requirements for recording, representing and replaying each of the different modes, and the research presented in this paper outlines some of the issues associated with this process.

The aim of this paper has been to begin to explore approaches that allow researchers simultaneously to review and analyse video, audio and textual records of naturally occurring communication. Such tools have the potential to provide a major resource for researchers in the field of applied linguistics and communication studies, film studies and drama in performance as well as in the field of face-to-face and remote human interaction. The development of research in this domain can also subsequently be extended to include pedagogic applications in the analysis of cross-cultural communication for modern foreign language learning as well as in professional discourse analysis, thus reinforcing the essentially interdisciplinary potential of applied research of which Michael Stubbs' work has been an exemplary instance.

Acknowledgements

The research on which this article is based is funded by the UK Economic and Social Research Council (ESRC), e-Social Science Research Node *DReSS* (www.ncess.ac.uk/nodes/digitalrecord), and the ESRC e-Social Science small

grants project *HeadTalk* (Grant N^o. RES-149-25-1016). Thanks are also due to Dawn Knight for providing copy for research reports which have in part at least formed the basis of this paper. All research reports so far are available via the National Centre for e-Social Science <http://www.ncess.ac.uk/research/sgp/headtalk/>.

The HeadTalk demonstrator, Cvision, is an interactive program which allows users to apply the visual tracking algorithm developed within the project to selected targets in an input video clip. Cvision takes as its input an avi format video file and produces a text file giving the estimated position of each target in each frame of that video. This may be imported into the current version of the DreSS 2 tool DRS. An output video may also be produced, if desired. This shows the results of tracking overlaid on the input video images and is a useful debugging and interpretation tool. Cvision allows multiple targets to be tracked in parallel, producing a description of the motion of each and showing intermediate results as they are obtained. Cvision is written in C++ and provided as a Windows .exe file via <http://www.ncess.ac.uk/research/sgp/headtalk/>. User documentation is also provided, and incorporates full details of the algorithm employed.

References

- Allwood, J., J. Nivre & E. Ahlsen 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics* 9: 1-26.
- Argyle, M. 1969. *Social Interaction*. London: Methuen.
- Baldry, A. and P. Thibault 2004. *Multimodal Transcription and Text Analysis* London: Equinox.
- Baldry, A. and P. Thibault 2006. Multimodal corpus linguistics, in *System and Corpus* (eds Thompson and Hunston) pp. 164-183, London: Equinox.
- Beattie, G. & H. Shovelton 2002. What properties of talk are associated with the generation of spontaneous iconic hand gestures? *British Journal of Social Psychology* 41, 403-417.
- Beattie, G.W. & H. Shovelton 1999. Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica* 123, 1/2: 1 – 30.
- Berger, K.W. & G.K. Popelka 1971. Extra-facial gestures in relation to speech-reading. *Journal of Communication Disorders* 3, 302 - 308.
- Baldry, A. and P. Thibault 2004. *Multimodal Transcription and Text Analysis* London: Equinox.
- Baldry, A. and P. Thibault 2006. "Multimodal corpus linguistics", in *System and Corpus* (eds Thompson and Hunston) pp.164-183, London: Equinox .
- Brown, R. 1986. *Social Psychology* (2nd ed.). New York: Free Press.
- Brundell, P. & D. Knight 2005. *Current Research and Tools to Support Data Intensive Analysis for Digital Records in e-Social Science*. Unpublished report. University of Nottingham.

- Crabtree, A., A. French, C. Greenhalgh, S. Benford, K. Cheverst, D. Fitton, M. Rouncefield and C. Graham 2005. "Developing digital records", DReSS Working Paper 2, e-Social Science Research Node: University of Nottingham.
- Ekman, P. & W. Friesen 1969. The repertoire of nonverbal behaviour: Categories, origins, usage and coding. *Semiotica* 1: 49-98.
- Ekman, P. & W. Friesen 1976. Measuring facial movement. *Journal of Nonverbal Behaviour* 1, 1.
- French, A., C. Greenhalgh, A. Crabtree, M. Wright, P. Brundell, A. Hampshire and T. Rodden 2006. Software Replay Tools for Time-based Social Science Data. Paper delivered at the 2nd annual international e-Social Science conference, June 2006, University of Manchester.
- Gardner, R. 1998. Between speaking and listening: The vocalisation of understandings. *Applied Linguistics*, 19, 204-224.
- Gardner, R. 2002. *When Listeners talk: Response tokens and listener stance*. Amsterdam: John Benjamins.
- Goldin-Meadow, S. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences* 3, 11: 419-429.
- Griffin, Z.M. 2004. The eyes are right when the mouth is wrong. *Psychological Science* 15, 12: 814.
- Gu, Y. 2006. Multimodal text analysis: A corpus linguistic approach to situated discourse. *Text and Talk* 26, 2: 127 – 167.
- Kapoor, A. & R.W. Picard 2001. A real-time head nod and shake detector. *ACM International Conference Proceedings Series*. 1-5.
- Kendon, A. 1980. Gesticulation and speech: Two aspects of the process of utterance. In M.R. Key (Ed.). *The Relationship of Verbal and Nonverbal communication*. The Hague: Mouton. pp. 207-227.
- Kendon, A. 1994. Do gestures communicate? A review. *Research on Language and Social Interaction* 27, 3: 175-200.
- Knight, D., S. Bayoumi, S. Mills, A. Crabtree, S. Adolphs, T. Pridmore & R. Carter 2006. Beyond the Text: Construction and Analysis of Multi-Modal Linguistic Corpora. Paper delivered at the 2nd annual international e-Social Science conference, June 2006, University of Manchester.
- Kress, G and T. van Leeuwen 1996. *Reading Images* Routledge, London.
- Lock, A. (ed.) *Action, gesture and symbol: the emergence of language*, London: Academic Press.
- Maynard, S.K. 1987. International functions of a nonverbal sign head movement in Japanese dyadic casual conversation. *Journal of Pragmatics* 11, 589-606.
- Maynard, S.K. 1990. Conversation management in contrast: listener response in Japanese and American English. *Journal of Pragmatics* 14, 397-412.
- McClave, E.Z. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics* 37, 7: 855-878.
- Nixon, M. and A. Aguado 2002. *Feature Extraction and Image Processing*, Oxford: Elsevier.

- O'Keeffe, A. & S. Adolphs forthcoming. Using a corpus to look at variational pragmatics: Response tokens in British and Irish discourse. in K.P. Schneider and A. Barron, ed., *Variational Pragmatics*. Amsterdam, Netherlands: John Benjamins.
- Rimé, B. & L. Schiaratura 1991. Gesture and speech. In Feldman, R. & Rimé, B. (eds.). *Fundamentals of nonverbal behaviour*. Cambridge: Cambridge University Press. pp. 239-281.
- Riseborough, M.G. 1981. Physiographic gestures as decoding facilitators: Three experiments exploring a neglected facet of communication. *Journal of Nonverbal behaviour* 5, 172 - 183.
- Rogers, W.T. 1978. The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances. *Human Communication Research*, 5, 54-62.
- 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.
- Schegloff, E. 1982. Discourse as interactional achievement: some uses of "uh huh" and other things that come between sentences. In D. Tannen (Ed.), *Analyzing discourse, text, and talk*. Washington, DC: Georgetown University Press. 71-93.
- Thompson, L.A. & D.W. Massaro 1986. Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of Experimental Child Psychology*, 42, 144-168.
- Wilcox, S. 2004. Language from gesture. *Behavioral and Brain Sciences* 27, 4: 525-526.
- Yngve, V. 1970. On getting a word in edgewise. In *Papers from the 6th Regional Meeting, Chicago Linguistic Society*. Chicago: Chicago Linguistic Society.