The Nottingham Multi-Modal Corpus: A Demonstration

Knight, D., Adolphs, S., Tennent, P. and Carter, R.
The University of Nottingham
The School of English Studies, The University of Nottingham, University Park, Nottingham, NG7 2RD, UK
E-mail: aexdk3@nottingham.ac.uk, Svenja.Adolphs@nottingham.ac.uk, pxt@Cs.Nott.AC.UK,
Ronald.Carter@nottingham.ac.uk

Abstract

This software demonstration overviews the developments made during the 3-year NCeSS funded *Understanding New Forms of the Digital Record for e-Social Science* project (DReSS) that was based at the University of Nottingham. The demo highlights the outcomes of a specific 'driver project' hosted by DReSS, which sought to combine the knowledge of linguists and the expertise of computer scientists in the construction of the multi-modal (MM hereafter) corpus software: the Digital Replay System (DRS). DRS presents 'data' in three different modes, as spoken (audio), video and textual records of real-life interactions, accurately aligning within a functional, searchable corpus setting (known as the Nottingham Multi-Modal Corpus: NMMC herein). The DRS environment therefore allows for the exploration of the lexical, prosodic and gestural features of conversation and how they interact in everyday speech. Further to this, the demonstration introduces a computer vision based gesture recognition system which has been constructed to allow for the detection and preliminary codification of gesture sequences. This gesture tracking system can be imported into DRS to enable an automated approach to the analysis of MM datasets.

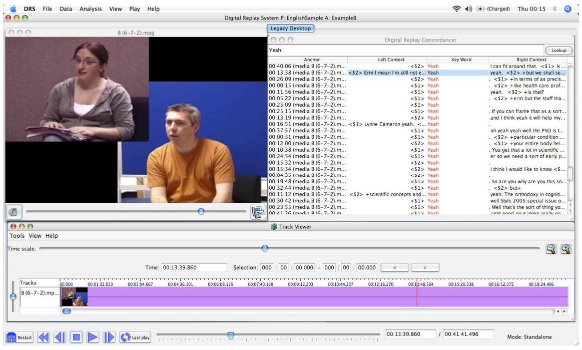
1. Introduction

This paper, and accompanying software demo, reports on some of the developments made to date on the 3-year ESRC (Economic and Social Research Council) funded DReSS (Understanding Digital Records for eSocial Science) interdisciplinary research project, based at the University of Nottingham. The linguistic concern of the project was to explore how we can utilise new textualities (MM datasets) in order to further develop the scope of Corpus Linguistic (CL hereafter) analysis. This paper discusses selected linguistic and technological procedures and requirements for developing such a MM corpus. We focus on the NMMC (Nottingham Multi-Modal Corpus, a 250,000 word corpus of single and dyadic conversational data taken from an academic discourse context), and we outline key practical issues that need to be explored in relation to mark-up and subsequent codification of linguistic and gesture phenomena.

2. Outlining the DRS

The Digital Replay System (DRS), the software used to interrogate the NMMC, aims to provide the linguist with the facility to display synchronised video, audio and textual data. In addition, perhaps most relevantly, it is integrated with a novel concordance tool which is capable of interrogating data constructed both from textual transcriptions anchored to video or audio and from coded annotations. Figure 1, below, shows an example of the concordance tool in use within the DRS environment. In this window, concordance lines for the search term *yeah* are displayed in the top right-hand panel and, as each concordance line is selected, the corresponding source video file is played to the left-hand side of the user-interface (UI hereafter).

For text based corpora (including current spoken corpora), concordance tools are nothing new. Wordsmith for example (http://www.lexically.net, see Scott, 1999) is a well known tool allowing an analyst to carry out concordance searches across large corpora of spoken or written discourse. It would be possible to export transcriptions from DRS to such a tool. However, by making such an export, we sacrifice many of the benefits of having a MM analysis tool such as DRS. DRS contains its own concordance tool. At its most basic level this allows the analyst to search across a transcription or collection of transcriptions (constituting a text only corpus) creating a concordance which displays the textual context of words or regular expressions.



Yeah

Anchor	Left Context	Key Word	Right Context
00:40:06 (media B (6-7-2).m...	<$2>	Yeah	I can fit around that. <$1> Is ...
00:13:38 (media B (6-7-2).m...	<$2> Erm I mean I'm still not e...	Yeah	yeah. <$2> +but we shall se...
00:26:09 (media B (6-7-2).m...	<$2>	Yeah	. <$1> +in terms of as precis...
00:00:15 (media B (6-7-2).m...	<$1>	Yeah	. <$2> +like health care prof...
00:11:56 (media B (6-7-2).m...	<$1>	Yeah	yeah. <$2> +is that?
00:05:22 (media B (6-7-2).m...	<$1>	Yeah	. <$2> +erm but the stuff tha...
00:25:09 (media B (6-7-2).m...	<$1>	Yeah	.
00:25:15 (media B (6-7-2).m...	<$1>	Yeah	. If you can frame that as a sort...
00:13:19 (media B (6-7-2).m...	<$2>	Yeah	and I think yeah it will help my...
00:16:51 (media B (6-7-2).m...	<$1> Lynne Cameron yeah. <...	Yeah	.
00:37:57 (media B (6-7-2).m...	<$1>	Yeah	oh yeah yeah well the PhD is t...
00:00:31 (media B (6-7-2).m...	<$1>	Yeah	. <$2> +particular condition ...
00:12:00 (media B (6-7-2).m...	<$2>	Yeah	. <$1> +your entire body hel...
00:10:38 (media B (6-7-2).m...	<$1>	Yeah	. You get that a lot in scientific ...
00:24:54 (media B (6-7-2).m...	<$1>	Yeah	er so we need a sort of early p...
00:15:32 (media B (6-7-2).m...	<$2>	Yeah	.
00:15:34 (media B (6-7-2).m...	<$2>	Yeah	I think I would like to know <$...
00:04:35 (media B (6-7-2).m...	<$2>	Yeah	.
00:19:48 (media B (6-7-2).m...	<$1>	Yeah	. So are you why are you this so...
00:32:44 (media B (6-7-2).m...	<$1>	Yeah	. <$2> but+
00:11:12 (media B (6-7-2).m...	<$2> +scientific concepts and...	Yeah	yeah. The orthodoxy in cogniti...
00:30:42 (media B (6-7-2).m...	<$1>	Yeah	well Style 2005 special issue o...
00:23:55 (media B (6-7-2).m...	<$1>	Yeah	. Well that's the sort of thing yo...
00:41:36 (media B (6-7-2).m...	<$1>	Yeah	right good no it looks really go...

Figure 1: The concordance tool in use within the DRS environment.

Perhaps the most immediate difference between a standard text-based corpus and a MM corpus is the need to use a timeline as a means of aligning all the different data streams. This may have originally been included as a logistical necessity, but in practice it allows a degree of flexibility that standard corpus software tools do not have. Further to this, it is important to note that the trend in corpus linguistics has been towards having all the data and metadata together in one file. DRS is much more flexible in this regard. Because it uses a timeline as an anchor, the user can attach as many transcripts or annotations to that timeline as is desired. This means that data and metadata can be stored in separate files, the text can be read easily by the user without being buried by hoards of metadata records, and vice versa.

Since such reference media can be organized, indexed and stored within the DRS, we can provide more than just a textual context. Simply clicking on an instance of the utterance, we can immediately display the video of that utterance occurring, providing a far greater degree of context than is available with more traditional text-only tools. In addition, we also have coded gestures as part of the NMMC, so the DRS concordancer allows the analyst to search across the codes as well, treating them in the same way as spoken utterances. The user can therefore use DRS as an analysis tool rather than just a read-only tool already provided by existing software, thus making DRS a useful interface for a wide variety of users.

It is important to note that the concordancer is still being enhanced in order to provide frequency counts of the data. The integration of this utility will eventually allow the linguist to research statistical or probabilistic characteristics of corpora, as well as to explore specific tokens, phrases and patterns of language usage (both verbal and non-verbal) in more detail. The current version of the DRS concordancer allows the analyst to search across texts as well as within texts, and provides a reference to the text from which specific concordances were derived. Once the tracker is integrated within DRS (see below), this feature can be used to allow the linguist to search for key terms and related 'tracked' gestures in order to start to map relationships between language and gesticulation.

This novel MM concordancer has led to the need for developing new approaches for coding and tagging language data, in order to align textual, video and audio data streams (see Adolphs and Carter, 2007 and Knight, 2006). Subsequently, this demo also reports on findings of explorations (using the concordance search facility) of relationships between the linguistic characteristics and context of specific gestures, and the physically descriptive representations of those gestures extracted from video data.

The study of this relationship leads to a greater understanding of the characteristics of verbal and non-verbal behaviour in natural conversation and the specific context of learning. This will allow us to explore in more detail the relationships between linguistic form and function in discourse, and how different, complex facets of meaning in discourse are constructed through the interplay of text, gesture and prosody (building on the seminal work of McNeill, 1992 and Kendon, 1990, 1994).

3. Coding using DRS

Codes in DRS are stored in a series of 'coding tracks'. Each of these tracks is based on a timeline and associated with a particular media file. Similarly, transcripts are stored as 'annotation tracks' which behave in the same way as coding tracks, though with free rather than structured annotations. Because each utterance or code has a time associated with it, as well as a reference media, it is possible to search across these different types looking for patterns with the original media instantly accessible in the correct place. This allows the analyst to examine the context of each artifact. In order to search the data effectively, a suitable tool is required.

DRS is equipped to support the annotation and coding of raw and semi-structured data through a multistage iterative process which includes "quick and dirty" qualitative exploration of the data. This is particularly useful where rapid accessing of data and rough annotation/coding is required in order to identify passages of interest and possible variables to be included in a coding scheme.

4. Annotating MM Corpora

Traditionally linguists have relied on text as a 'point of entry' for corpus research. However, one of the fundamental aims of this project is that all modes should be equally accessible to corpus searches, allowing not only text-based linguists but also researchers investigating the use of gesture to access data.

This principle has led to the need for new approaches for annotating and coding textual language data, in order to align them with video and audio data streams, thus enabling subsequent analysis (see Adolphs & Carter, 2007; Knight, 2006). For a MM corpus to be of use to the broader research community all streams should be accessible in order to facilitate research.

Current annotation schemes that are equipped for both gesture and speech (including, but not limited to those used within the field of linguistics) tend to only look at each mode in turn, as Baldry and Thibault (2006: 148) emphasise:

> 'In spite of the important advances made in the past 30 or so years in the development of linguistic corpora and related techniques of analysis, a central and unexamined theoretical problem remains, namely that the methods adapted for collecting and coding texts isolate the linguistic semiotic from the other semiotic modalities with which language interacts…. [In] other words, linguistic corpora as so far conceived remains intra-semiotic in orientation…. [In] contrast MM corpora are, by definition, inter-semiotic in their analytical procedures and theoretical orientations.'

Many schemes do exist, however, which depict the basic semiotic relationship between verbalisations and gesture (early coding schemes of this nature are provided by Efron 1941 and Ekman and Friesen 1968, 1969). These mark-up the occasions where gestures co-occur (or not) with the speech, and state whether the basic discoursal function of the gestures and speech 'overlap', are 'disjunct' and so on, or if the concurrent verbalisation or gesture is more 'specific' than the other sign at a given moment (for more details see Evans et al., 2001: 316). These schemes may be a useful starting point for labeling information in each mode, which can be further supplemented to cater for the semantic properties of individual features.

An example of a coding scheme that deals with defining a range of gestures based upon sequences of kinesic movements (that occur during speech) was been drawn up by Frey et al. (1983). Other more detailed kinesic coding schemes exist which attempt to define more explicitly the specific action, size, shape and relative position of movements throughout gesticulation (see Holler and Beattie, 2002, 2003, 2004; McNeill, 1985, 1992; Ekman & Friesen 1968, 1969). However, these schemes are limited in their utility for marking up the linguistic function of such sequences, and their explicit relationship to spoken discourse Other available coding schemes are not designed to provide the tools for more pragmatic analyses of language, nor to facilitate the integration of analyses of non-verbal and verbal behaviour as interrelated channels for expressing and receiving messages in discourse.

Current schemes that do classify the verbal and the visual only tend to deal with the typological features of MM talk. An example of this is given by Cerrato (2004: 26, also see Holler & Beattie's 'binary coding scheme for iconic gestures', 2002) who marks up a range HH and HCI conversations according to, primarily, whether it is a word (marked as W), phrase (marked as P), sentences (marked as S) and gestures (marked as G). Indeed steps to facilitate the exploration of both modes in conjunction have been made by various researchers and research teams (for example Cerrato 2004, discussed in more detail in chapter 4, and Dybkjær & Ole Bernsen, 2004).

Other key limitations with current coding and annotation schemes and tools are that they are not always available for general use. Instead they are often designed to meet the address a particular research question and so are difficult to expand beyond the remit of their associated research projects. For example, more extensive coding schemes that are equipped for dealing with both gesture and speech (a variety of schemes are discussed at length by Church & Goldin-Meadow, 1986 and Bavelas, 1994) are generally designed primarily to model sign language and facial expressions specifically (which can also be used for determining mouth movements in speech, as is common in HCI studies). Examples of such coding schemes are the HamNoSys (Hamburg Notation System, see Prillwitz et al., 1989), the MPI GesturePhone (from the Max Planck Institute; which transcribes signs as speech) as well the MPI Movement Phase Coding Scheme which is designed to code gestures and signs which co-occur with talk (Kita et al., 1997).

The coding scheme that is perhaps closest to the requirements of MM corpora is the MPI Movement Phase Coding Scheme. This is described as 'a syntagmatic rule system for movement phases that applies to both co-speech gestures and signs' (Knudsen et al., 2002). However, this system does not provide detailed codes for the functional significance of the different characteristics of talk. It is a scheme that was developed at the MPI in order to allow for the referencing of video files. Annotations made with this scheme can be conducted using another MPI tool, MediaTagger, and are input into the software EUDICO for further analysis and the representation of data.

The development of a coding system that is more transferable across the different data streams in the MM corpus would be useful for the purpose of linguistic analysis. It would allow us to connect the pragmatic and semantic properties of the two gesture and speech and enable cross referencing between the two. This would make it easier to search for patterns in concordances of the data in order to explore the interplay between language and gesture in the generation of meaning.

Despite this the ISLE project started to make steps towards outlining the foundational requirements for creating 'International Standards for Language Engineering' (Dybkjær & Ole Bernsen, 2004: 1). Such set standards may be of use to the development of MM corpora. These standards are known specifically as NIMMs; Natural Interaction and MM Annotation Schemes. ISLE is a project that is based on the notion that there is a need for a 'coding scheme of a general purpose' to be constructed to deal with the 'cross-level and cross modality coding' of naturally occurring language data (Dybkjær & Ole Bernsen, 2004: 2-3, also refer to Wittenburg et al., 2000).

However, since gesticulations are so complex and variable in nature, it would appear difficult to create a comprehensive scheme for annotating *every* feature of gesture-in-use. Depending on the perspective of research, gestures may also be seen to have different semantic or discursive functions in the discourse. Thus it would be difficult to mark-up each respective function.

5. Coding the NMMC

Despite practical constraints, we have aimed to encode the NMMC data in a way that will 'allow for the maximum usability and reusability of encoded texts' (Ide, 1998: 1). The basic coding rubric adopted can be seen in figure 2. In order to explore the 'interaction of language and gesture-in-use for the generation of meaning in discourse', which has been the central aim for linguistic analyses using the NMMC, we initially focused upon classifying movement features and linguistic features independently.

Gestures are classified in a top-down fashion, with the analyst working to firstly define the specific form of gesture, before proceeding to establish the linguistic function (pragmatic category) of such. Knowledge from the tracking output and manual analyses determine the shape and direction of hands and whether one or both of the hands are moving at any one point. . These are then classified using the typological descriptions of gestures that are available in gesture research. The aim is to establish whether the movement features of the gesture best attribute it to being *Iconic*, *Metaphoric*, *Beat-like*, *Cohesive* or *Deictic* in nature (see McNeill, 1995, 1985, similar paradigms are seen in McNeill et al., 1994: 224; Richmond et al., 1991: 57; Kendon, 1994).

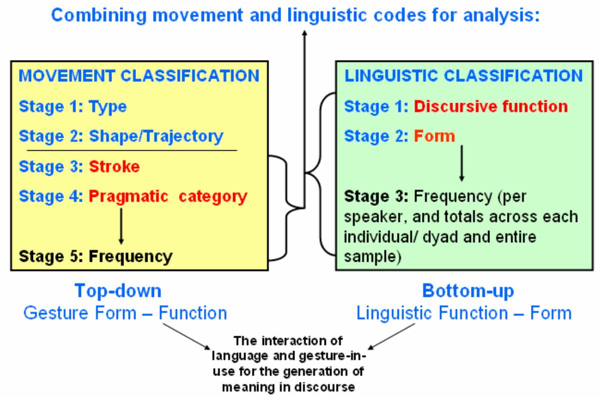

Combining movement and linguistic codes for analysis:
MOVEMENT CLASSIFICATION
Stage 1: Type
Stage 2: Shape/Trajectory
Stage 3: Stroke
Stage 4: Pragmatic category
Stage 5: Frequency
LINGUISTIC CLASSIFICATION
Stage 1: Discursive function
Stage 2: Form
Stage 3: Frequency (per speaker, and totals across each individual/ dyad and entire sample)
Top-down
Gesture Form – Function
Bottom-up
Linguistic Function – Form
The interaction of language and gesture-in-use for the generation of meaning in discourse

Figure 2: Coding verbal and non-verbal features of talk

In instances whether gestures co-occur specifically with speaker verbalisations (rather than with recipient gesticulations), we are working with a separate classification system in a bottom-up manner, exploring first the discursive function of co-occurring text before looking in more detail at specific tokens and phrases (which are separately encoded, see Knight and Adolphs, 2008 for details). As a final measure this information is combined in order to explore more closely specific words or phrases that are likely to co-occur with specific gestures throughout the gesture phase (and at the *stroke;* the most emphatic point, in particular).

6. Analysing gesture in the NMMC

In addition to the NMMC DRS interface, a further aim for DReSS was to develop tools which model gesture-in-talk, with the ability to monitor the *function*, *timing* and *response* (if any) of all participants, to gain an increased understanding of their role in discourse.

Although it may be feasible to manually extract and observe specific sequences of gesticulation as they occur in 10 minutes or even 5 hours of video data, it should be acknowledged that an increase in length of video data makes this method less practical and cost-effective to use. In the same way that the manual examination of pre-electronic corpora was time-consuming and error prone, the manual strategies presented here are not yet automated and rely on manual analysis The linguist has to trawl through each second of data to find features of interest, before manually marking up and encoding those features, and manipulating them before patterns and general observations can be explored.

It may therefore be appropriate to exploit a more automatic digital approach for such analysis in future. This should detect and ultimately define and encode gesture-in-talk (based on parameters pre-determined by the analyst) at high speed, and thus reduce the amount of time required to undertake such operations. In addition to this, automated methods should help to provide scientifically verifiable parameters of gesture categories and codes. One such 'automatic' approach has been developed by Computer Vision experts at the University of Nottingham (and has been tested as part of the DReSS project) in the form of a 2D gesture algorithm, which can be seen in figures 3 and 4 (for information on the technological specifications of the algorithm see Knight et al., 2006 and Evans & Naeem, 2007).

The tracker is applied to a video (represented in the form of circular nodes) of a speaker and reports in each frame the position of, for example, the speaker's hands in relation to his torso. These targets, which can be adjusted in terms of size in relation to the image, are manually positioned at the start of the video and subsequently, as the tracking is initiated, we are presented with three vertically positioned lines marking four zones on the image, R1 to R4 (R2 and R3 mark the area within shoulder width of the participant, acting as a perceived natural resting point for the arms, hence R1 and R4 mark regions beyond shoulder width).

The algorithm tracks the video denoting in which region the left hand (labeled as **R** by the tracker, since it is located to the right of the video image) and right hand (labeled as **L** by the tracker, since it is located to the left of the video image) are located in each frame. So as the video is played movement of each hand is denoted by changes in the x-axis position of R and L across the boundaries of these vertical lines. Figure 4 (overleaf) shows an alternative location matrix that can used with the tracker, dividing the video image into 16 separate zones (based on McNeill's diagram for gesture space encoding, 1992: 378) for a more detailed account of specific the horizontal and vertical movements of each hand.

The movement of each hand can therefore be denoted as a change in x-axis based region location of the hand. So when using the tracker seen in figure 3 (overleaf), we see a sequence of outputted zone 3 for frames 1 to 7, which changes to a sequence of zone 4 for frames 8 to 16 for **R**, this notifies the analyst that the left hand has moved across one zone boundary to the right during these frames. In theory, in order to track larger hand movements, the analyst can pre-determine a specific sequence of movements which can be searched and coded in the output data. So if, for example, the analyst had an interest in exploring a specific pattern of movement, considered to be of an *iconic* nature, i.e. a specific combination of the spontaneous hand movements which complement or somehow enhance the semantic information conveyed within a conversation, it would be possible to use the hand tracker to facilitate the definition of such gestures across the corpus (for in-depth discussions on iconics and other forms of gesticulation, see studies by Ekman and Friesen, 1969; Kendon, 1972, 1980, 1982, 1983; Argyle, 1975; McNeill, 1985, 1992; Chalwa and Krauss, 1994 and Beattie and Shovelton, 2002).

In both cases the tracker outputs 'raw' data into the Excel spreadsheet consisting of a frame-by-frame account of the region location of each hand (in terms of it's position within the numbered matrix; comprising of a sequence of numbers for each frame for **R** and **L**). The movement of each hand is therefore denoted as a change in region location of the hand, so for example for **R** hand (the left hand), we see a sequence of outputted zone 3 for frames 1 to 7, which changes to a sequence of zone 4 for frames 8 to 16. Ergo this notifies the analyst that the **R** hand has moved across one zone boundary to the right during these frames. Using this output the analyst would be required to 'teach' the tracking system be means of pre-defining the combination of movements to be coded as 'iconic gesture 1', for example (so perhaps a sequence of **R** or **L** hand movements into from R1 to R4 and back to R1 across x amounts of frames, for the tracker seen in figure 3), in order to convert the raw output into data which is useable.

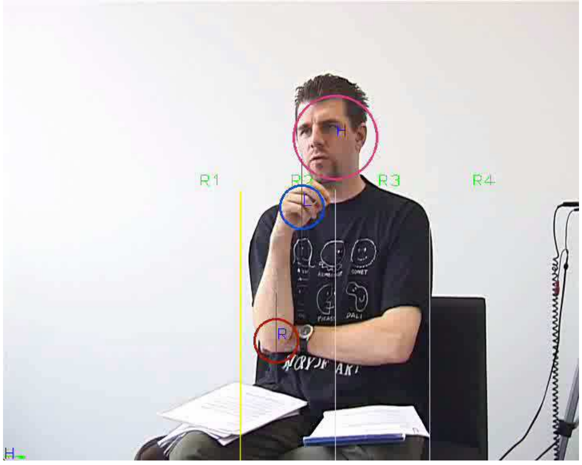


Figure 3: The initial 4 regions of the Hand Tracker

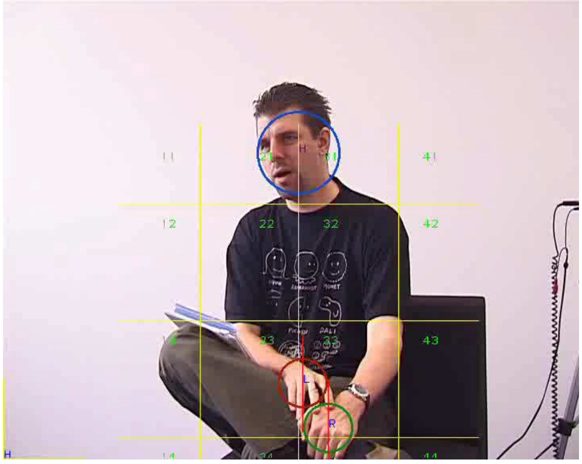


Figure 4: A 16 region version of the Hand Tracker

The raw data can, however, be plotted on to a basic graph, as seen in figure 5, which as a basic measure, informs the analyst whether movement does or does not occur at points throughout the video (this plot can be integrated into the DRS software). The graph maps the movement of the **L** and **R** hand across each region on the movement matrix, thus denoting movements which occur in a left or right location. This notion of movement Vs no-movement acts a useful preliminary step to classifying and encoding specific movement sequences, one which can be enacted automatically, again decreasing the amount of time required to manually extract such information. However, further to this the analyst is obviously required to determine whether these movements are in fact examples of gesticulation rather than fidgeting, for example, as regardless of how sensitive the system is, the complex nature of bodily movement makes it near impossible to determine such a difference fully automatically.

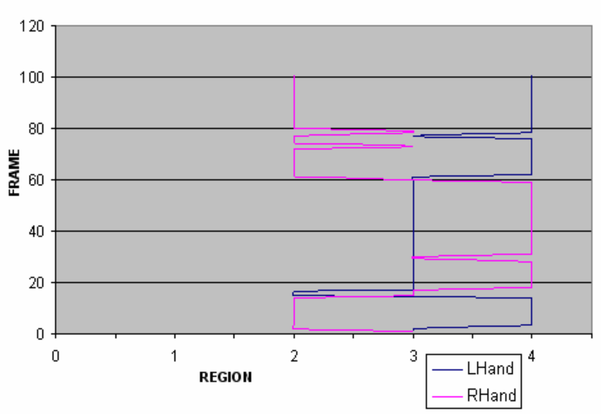


Figure 5: Plotting the tracking information (using the initial 4 region information, seen in figure 3)

It is important to note that the tracker is designed to allow the analyst to track more than one image in the same frame at the same time. In other words it has the ability for the user to apply the tracker on pre-recorded, digitised images which in theory can include up to two participants in each recorded image frame, so both participants as recorded in the NMMC corpus data comprising of dyadic academic supervisions. However, after extensive testing, it was discovered that the tracker appears to be at its most effective when the video is of high quality (.avi) with a high resolution, with the image of each participant shown as close-up and large scale as possible. This is because smaller, lower quality images were more likely to lose the tracking target locations instantly. This requirement proved to be slightly problematic to adhere to when dealing with the streamed two-party videos from the supervision sessions because the reduction in the physical size and associated quality of the image seen in such aligned videos causes the tracker to readily lose the target locations, making it difficult for the CV algorithm to adequately track these locations. In such situations it was found that even when frequent *debugging* (when the tracker loses it's desired targets and is thus manually stopped by the analyst and the target features are redefined and relocated before the tracking is resumed) was undertaken, target locations were often instantly lost when tracking recommenced.

We further attempted to run the tracker off the original, individual .avi videos from each recording and found that, in general, the tracking algorithm was able to track the desired bodily locations with increased levels of consistency (i.e. with decreased amounts of debugging required) and accuracy with such data. Using the individual source videos rather than those which have been aligned makes the process of tracking even more lengthy as each individual participant needs to be tracked in turn rather than simultaneously. However, using the split screen version of the videos, it can be even more difficult to watch both images simultaneously and accurately stop and debug the tracker as required. Consequently, it was deemed more beneficial, and in the long term more accurate, to deal with each image individually, before attempting to align results at a later date.

Kapoor & Picard point out, in their development of a 'real-time detection' and classification tool, even with salient gestures, for example head nods and shakes, it is difficult to fully automate this stage (2001). This is due to the fact that gestures-in-talk are spontaneous, idiosyncratic (Kendon, 1992) and transient (Bavelas, 1994: 209), and are generally seen to contain no standard forms in conversation (it is unlikely that two hand motions will be exactly the same, for example). Instead, they differ according to user, intensity, meaning and in terms of how the head or hand is rotated, i.e. whether it is simply a rigid up and down, left or right movement, or whether there is more of an up and slight rotation. This complexity in form means it is not easy to accurately encode and quantify particular movement features, especially if relying on purely automated methods. The intervention of the expert human analyst is still paramount in this context.

The gesture tracker, in its current, generates a fairly simplistic set of codes. It is generally possible to tell if a large gesture has occurred, but difficult to differentiate between different types of gesture. To some extent it serves to create a 'code-template' from which a skilled analyst can apply a more detailed coding scheme to generate a more complete description of the gestures captured in a given video session. Even from the tracker's simplistic codes it is possible to search for co-occurrences of gesture and utterance, but with a more detailed coding track – generated either by simply hand-coding the video using DRS's comprehensive coding tools, or by taking the code-template generated by the tracker and filling out the detail.

7. Summary

This demonstration paper has started to outline some of the technical and practical problems and considerations faced in the development and exploration of MM corpora. It presents a novel MM corpus UI (user-interface), the DRS. DRS provides the analyst with an easy-to-use corpus tool-bench for the exploration of relationships between the linguistic characteristics and context of specific gestures, and the physically descriptive representations of those gestures extracted from video data (using the novel MM concordancer).

8. Acknowledgements

The research discussed on which this article is based is funded by the UK Economic and Social Research Council (ESRC), e-Social Science Research Node *DReSS* (Grant N$^{o.}$ RES-149-25-0035, http://www.ncess.ac.uk/research/digital_records), and the ESRC e-Social Science small grants project *HeadTalk* (Grant N$^{o.}$ RES-149-25-1016).

9. References

Adolphs, S. & Carter, R. (2007). Beyond the word: New challenges in analysing corpora of spoken English. *European Journal of English Studies 11*(2).

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

Baldry, A. & Thibault, P.J. (2006). *Multimodal Transcription and Text Analysis: A multimedia toolkit and coursebook.* London: Equinox.

Bavelas, J.B. (1994). Gestures as part of speech: methodological implications. *Research on Language and Social Interaction 27,* 3: 201-221.

Beattie, G. & Shovelton, H. (2002). What properties of talk are associated with the generation of spontaneous iconic hand gestures? *British Journal of Social Psychology 41*, 3: 403-417.

Cerrato, L. (2004). A coding scheme for the annotation of feedback phenomenon in everyday speech. *LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, Lisboa. pp. 25-28.

Chawla, P. & Krauss, R. M. (1994). Gesture and speech in spontaneous and rehearsed narratives. *Journal of Experimental Social Psychology 30*: 580-601.

Church, R.B. & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition 23*, 1: 43-71.

Dybkjær, L. & Ole Bernsen, N. (2004) Recommendations for natural interactivity and multimodal annotation schemes. *Proceedings of the LREC'2004 Workshop on Multimodal Corpora*, Lisbon, Portugal. pp. 5-8.

Efron, D. 1972. (1941). *Gesture, Race and Culture.* The Hague: Mouton & Co.

Ekman, P. & Friesen, W. (1968). Nonverbal behavior in psychotherapy research. In J. Shlien (ed), *Research in Psychotherapy*. Vol. III. American Psychological Association. pp.179-216.

Ekman, P. & Friesen, W. (1969). The repertoire of non-verbal behavior: Categories, origins, usage and coding. *Semiotica 1*, 1: 49-98.

Evans, J.L., Alibali, M.W. & McNeill, N.M. (2001). Divergence of verbal expression and embodied knowledge: Evidence from speech and gesture in children with specific language impairment. *Language and Cognitive Processes 16*, 2-3: 309-331.

Evans, D. and Naeem, A. (2007). "Using visual tracking to link text and gesture in studies of natural discourse", *Online Proceedings of the Cross Disciplinary Research Group Conference 'Exploring Avenues to Cross-Disciplinary Research'*, November 7, University of Nottingham.

Frey, S., Hirsbrunner, H.P., Florin, A., Daw, W. & Crawford, R. (1983). A unified approach to the investigation of nonverbal and verbal behaviour in communication research. In Doise, W. & Moscovici, S. (Eds.), *Current issues in European Social*

Psychology. Cambridge: Cambridge University Press.

Holler, J. & Beattie, G. (2002). A micro-analytic investigation of how iconic gestures and speech represent core semantic features in talk. *Semiotica 142*, 1-4: 31-69.

Holler, J. & Beattie, G. (2003). How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? *Semiotica 146*, 1-4: 81-116.

Holler, J. & Beattie, G.W. (2004). The interaction of iconic gesture and speech. *5th International Gesture Workshop*, Genova, Italy. Selected Revised Papers. Heidelberg: Springer Verlag.

Ide, N. (1998). Corpus encoding standard: SGML guidelines for encoding linguistic corpora. *First International Language Resources and Evaluation Conference*, Granada, Spain.

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

Kendon, A. (1972). Some relationships between body motion and speech. In Seigman, A. & Pope, B. (Eds.), *Studies in Dyadic Communication*. Elmsford, New York: Pergamon Press. pp.177-216.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In Key, M.R. (Ed), *The Relation between Verbal and Non-Verbal Communication*. pp. 207-227.

Kendon, A. (1982). The organisation of behaviour in face-to-face interaction: observations on the development of a methodology. In Scherer, K.R. & Ekman, P. (eds) *Handbook of Methods in Nonverbal Behaviour Research*. Cambridge: Cambridge University Press.

Kendon, A. (1983). Gesture and Speech: How they interact. In Wiemann, J. & Harrison, R. (Eds.), *Nonverbal Interaction*. California: Sage Publications. pp.13-46.

Kendon, A. (1990) *Conducting Interaction*. Cambridge: Cambridge University Press.

Kendon, A. (1992). Some recent work from Italy on quotable gestures ('emblems'). *Journal of Linguistic Anthropology 2*, 1: 77-93.

Kendon, A. (1994). Do gestures communicate? A review. *Research on Language and Social Interaction 27*, 3: 175-200.

Kita, S., van Gijn, I., & van der Hulst, H. (1997). Movement Phase in Signs and Co-Speech Gestures, and Their Transcriptions by Human Coders. *Gesture Workshop 1997*: 23-35.

Knight, D. (2006). ‘Corpora: The Next Generation’, Part of the AHRC funded online *Introduction to Corpus Investigative Techniques*, The University of Birmingham. http://www.humcorp.bham.ac.uk/

Knight, D. and Adolphs, S. (In Press, 2008) Multi-modal corpus pragmatics: the case of active listenership. In Romeo, J. (ed.) *Corpus and Pragmatics*. Berlin and New York: Mouton de Gruyter.

Knight, D., Bayoumi, S., Mills, S., Crabtree, A., Adolphs, S., Pridmore, T. & Carter, R. (2006). Beyond the Text: Construction and Analysis of Multi-Modal Linguistic Corpora. Published in the *Proceedings of the 2nd International Conference on e-Social Science, Manchester, 28 - 30 June 2006*.

Knudsen, M. W., Martin, J.-C., Dybkjær, L., Ayuso, M. J. M, N., Bernsen, N. O., Carletta, J., Kita, S., Heid, U., Llisterri, J., Pelachaud, C., Poggi, I., Reithinger, N., van ElsWijk, G. & Wittenburg, P. (2002). Survey of Multimodal Annotation Schemes and Best Practice. *ISLE Deliverable D9.1, 2002*.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review 92*, 3: 350-371.

McNeill, D. (1992). *Hand and Mind*. Chicago: The University of Chicago Press.

McNeill, D. (1995). *Hand and mind: What gestures reveal about thought*. Chicago: The University of Chicago Press.

McNeill, D., Cassell, J., McCullough, K-E. (1994). Communicative effects of speech-mismatches gestures. *Research on Language and Social Interaction 27*, 3: 223-237.

Prillwitz, S., Leven, R., Zienert, H., Hanke, T. & Henning, J. (1989). *HamNoSys. Version 2.0. Hamburg Notation System for Sign Language. An Introductory Guide*. Hamburg: Signum.

Richmond, V.P., McCroskey, J.C. & Payne, S.K. (1991). *Nonverbal Behaviour in Interpersonal Relations*. Prentice Hall: New Jersey.

Scott, M. (1999). *Wordsmith Tools*. Oxford: Oxford University Press.

Wittenburg, P., Broeder, D. & Sloman, B. (2000). Meta-description for language resources. *EAGLES/ ISLE White paper.* Available online from http://www.mpi.nl/world/ISLE/documents/papers/white_paper_11.pdf. [Accessed 2006-02-10]