

Beyond the Text: Construction and Analysis of Multi-Modal Linguistic Corpora

Dawn Knight, Sahar Bayoumi, Steve Mills, Andy Crabtree, Svenja Adolphs, Tony Pridmore, Ronald Carter

This paper addresses some of the linguistic and technological procedures and requirements of the next generation of tools for the analysis of spoken linguistic corpora. It reports on preliminary developments of an ESRC funded interdisciplinary project at the University of Nottingham. It specifically focuses on key methodological and technical issues related to the mark-up, coding and representation of multi-modal communication data. The paper builds on traditional approaches in the area of spoken corpus linguistics, which use multi-million word databases of textual renderings of naturally occurring conversation with the aim of identifying recurring patterns of lexical and grammatical patterns within sequences of interactions. The identification, classification and representation of accompanying gestural elements are explored in relation to the expression of active listenership.

Introducing Corpus Linguistics

Corpus linguistics (CL) is best understood as a methodology that can be used in different disciplines. It is an empirically based approach which involves the processing of language data based on large textual databases and the subsequent interpretation of the output. In linguistics the term corpus is used to describe a 'large and principled collection of natural texts' (Biber et al 1998:4). This collection of texts allows us to provide factual observations about language use across various different contexts both on a micro scale, within or across sentences, and a wider scale denoting differences across texts and subjects. The analysis of corpora, therefore, allows us to give new and different perspectives to language description and as a result there are many potential uses of corpora, including English Language Teaching and lexicography.

While most research in the area of corpus linguistics to date has focused on written corpora, there is an increasing interest in the exploration of spoken data with the use of corpus linguistic techniques. Various spoken corpora have been developed over the past two decades which in turn have led to new descriptions of the way in which we use lexis and grammar in different types of spoken contexts (see Carter and McCarthy, 2006).

However, one of the issues with spoken corpora is that they currently only provide a representation of the spoken features of discourse that can be easily rendered into a standardised textual format. This is problematic as, in essence, communication does not rely upon the spoken word alone. This notion was first explored by Dobrogaev (1929, reported in Kendon, 1980: 225) who conducted a study of communication, requesting subjects to participate in dyadic conversations whilst suppressing head, body and hand movements, a task which was found to be next to impossible to achieve. This was because, as is now widely recognised, real, naturalistic communication relies upon multiple modes of expression, a combination of the verbal and the nonverbal, 'expressive signs, signals and cues' (also known as 'paralinguistic' features, Haiman, 1998:23). Communication may therefore be best

described as operating within a complex network of direct and indirect ‘semiotic channels’ (Brown, 1986: 409) including, for example, patterns of intonation and vocabulary use as well as gesture and facial expression.

Brown further developed this idea by describing communication as a complex network composed of various direct and indirect ‘semiotic channels’ (1986: 409). These channels include, for example, patterns of intonation- aspects of voice, language used as well as facial expressions and gestures made throughout the course of the interaction. Brown highlighted that although these different channels essentially work independently, they also operate simultaneously during communication and regularly complement as well as regulate each other (1986: 409). Current corpus linguistics methods inhibit our abilities to explore these features of discourse, as they are only able to present textual renderings and records of interaction, failing to represent discourse *beyond the text*.

Corpora: The next generation

In order to develop the scope of the kind of evidence we can extract from a spoken corpus, a multi-modal approach needs to be developed, providing the tools for exploring discourse beyond the text in order to look at the verbal and the non-verbal elements simultaneously in specific contexts of communication. This will allow us to explore the relationship between the two, assessing both individually and collectively their role in conversational exchanges and the development and expression of meaning in discourse.

To this end we are exploring the requirements for the development of *multi-modal corpora*, using video data recorded from conversational exchanges that is to be *streamed* with the verbal data. In order to achieve this, we explore different technological procedures for the collection of the multi-modes of data in order to develop an integrated way of annotating different elements of communicative events.

This paper reports on some preliminary developments of an ESRC funded interdisciplinary demonstrator project (HeadTalk) at the University of Nottingham. This project seeks to provide a rubric for the development of a multi-modal, multi-media corpus tool which can be utilised to explore gesture-in-talk in more detail, thus extending the utility of current textual corpora.

The HeadTalk Project

The aim of our project is to provide resources for developing our understanding of and research into the characteristics of a specific ‘semiotic channel’; that of head nods. Head nods are recognised as one of the most salient gestures in communication, and are understood as a type of ‘backchannel’, a term first coined by Yngve, in a study of turn taking in conversation. He describes backchannels as channels ‘over which the person who has the turn receives short messages such as *yes* and *uh-huh* without relinquishing the turn’ (Yngve, 1970: 568, see also Roger & Nesshover, 1987). Essentially, backchannels exist as ‘the antithesis of interruptions’, responses to a stimulus which are not intended to take ‘control of the floor’, as a complete turn would, but are intended to offer some form of relevant feedback to the speaker (Mott & Petrie, 1995).¹

There are a number of different forms that backchannels can take. In terms of vocalised, verbal and non-verbal backchannels, Gardner (1997: 18) identifies the following forms: (see also Maynard, 1989, 1990, 1997, Schegloff, 1982 & Gardner, 1998, 2002)

- Minimal vocalised acknowledgements
- Brief agreements
- News-marking items
- Appreciative, sympathetic and evaluative items
- Clarification requests (at certain points- as repairs)
- Laughter, sighs and other sympathetic vocalisations.

Head nods exist as part of a separate group of backchannels, consisting of non-vocalised kinesic signals and proxemic movement. They are often viewed as one of the most highly conventionalised, salient gestures of communication and although not spoken, such gestures are integral in determining the meaning and functions of linguistic components and can also act as a means for hearers to register and evaluate what is being said. Yet, although on a basic level head nods adopt the same highly conventionalised, and to some extent, easily definable *form*, that of up and down motion of the head, their given relevance or meaning in discourse is perhaps not as straightforward as they do not have just one specific *function*.

Indeed, the use of head nods in conversation extends beyond the most common connotations of affirmation and negation (as highlighted in McClave, 2000: 859). Head nods are also deictic and are 'directly related to the discourse structure of an utterance' (Kendon, 1972: 195), and therefore they are vital for conversational maintenance and management. It is essential to explore the function of a given backchannel in conjunction with its form in order to gain an accurate definition of the phenomenon. This is because it is difficult to accurately distinguish between backchannels and turns in conversation (see Duncan & Neiderehe, 1974 & Tottie, 1991: 260). Indeed, an ability to define backchannels and distinguish them from turns is needed by participants throughout a conversational exchange, since participants need to *detect* and *analyse* them in order to *understand* the significance of each of the backchannels' relevant functions.

When successful therefore, gestures such as head nods can act as an effective 'substitute for speech' (Goldin-Meadow, 1999: 419). However such saliency does not always exist and as a result misunderstanding can occur, impeding the effectiveness of the communication. Misunderstandings can also be due to a variety of factors. These include a lack of 'substantive information' held by the recipient of the message, a cultural difference, intentional or unintentional, 'mismatches' between gestures and spoken messages (Goldin-Meadow, 1999: 426). 'Mismatches' also occur because the meaning of head nods differ depending upon a range of complex factors, composing a 'gestural complex scenario' wherein 'human language is a highly specialized, evolutionary manifestation of a multimodal gestural complex' (Wilcox, 2004: 256).

In addition to this, as explored in the notion of the semiotic system, the verbal and non-verbal channels of the discourse 'complement and regulate each other' (Brown, 1986: 409), yet do not necessarily do so in a consistent way throughout the conversation, even with the same participants. So, for example, at a given point a head nod may be used in conjunction with the *yeah* as the mark of agreement or a convergence token, but elsewhere the head nod alone will mark the convergence, without an accompanying vocalisation. Conversely, a different gesture used in conjunction with the same utterance does not always derive a different meaning within a specific conversation.

The relevance of backchannels are also specific to the subjects involved and the topics of conversation and since, like gestures in general, head nods are understood as culturally dependant 'emblems' (McClave, 2000: 860, adapted from Efron, 1941 & Ekman & Freisen,

1969), they are also stringently controlled by aspects of context, such as time, place and culture. In short, backchannels are specific to a channel, use and function at a specific time and place and therefore are not always transferable across similar socio-contextual situations, making their detection, exploration and analysis difficult.

So in order to gain a better understanding of head nods, it is important to develop our abilities to 'read' them in terms of their given function, when and where they occur in the co-text and the context, that is in terms of what forms come before and after them in order to provide the foundations of a more complete understanding of discourse (Goldin-Meadow, 1999: 425). Thus a tool which seeks to model head nods (as proposed here) requires the ability to monitor the function, timing, significance and response (if any) of all parties involved, in order to gain an increased understanding of their significance. It is difficult to create accurate tools for identifying and labelling backchannels as the complexities and inconsistencies of backchannel use in natural conversation makes it difficult to explore the phenomena in great detail.

Coding Backchannels: A linguistic approach

One of the key areas of concern of this project is how the headnods in our data will be encoded. In terms of verbal realisations of backchannels most existing schemes focus upon grouping these in terms of their function, in other words their basic roles in discourse. This is a useful point of categorisation as every backchannel has a function in discourse, even if it is unconscious to the interlocutor of the term. Indeed a wealth of research exists which agrees that 'backchannels have more than one macro function' (O'Keeffe and Adolphs forthcoming, 2006) as defined below (see also Schegloff, 1989 & Maynard, 1989). As a guide to the key functions we are using the rubric provided by O'Keeffe and Adolphs, which identifies the following:

- **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 backchannels, the question of how verbal and visual realisations interact within and across such categories has remained largely under-explored.

Both verbal and non-verbal backchannels have the potential to vary according to their placement (i.e. the specific point(s) in discourse in which they occur), intensity and duration, and we therefore need to identify 'shared' groupings or categories which take into account these multi-modal characteristics. In order to develop an understanding of the nature of such shared groupings, we explore different techniques for detecting, coding and replaying visual and verbal elements of backchannels in a multi-modal corpus.

Automated Analysis of Conversational Gesture

Any tools capable of supporting the construction and analysis of video-based multi-modal corpora must be underpinned by computer vision techniques. The goal of computer vision is to extract meaningful information about the real world from an image, set or sequence of

images. Early work focused on the recovery of static properties of viewed objects, such as surface shape and colour, and two-dimensional motion estimates describing the movement of image features across the image plane. As the field has matured attention has steadily moved towards higher-level interpretation. The recognition of domain-specific events, including gestures made by human participants in the course of everyday activities, is now an active topic with the computer vision community.

In the context of multi-modal corpora, the goal of Headtalk is to identify computer vision techniques that may be used to extract descriptions of gesture from video recordings of natural conversation. These gesture descriptions, along with the more conventional transcripts and annotations, provide the basis for construction and analysis of multi-modal corpora.

The largest application area for visual gesture recognition is currently the analysis of sign language (Ong and Ranganath 2005). Sign language gestures are largely defined by the shape, orientation, location, and motion of the hands, but meaning may be modified by other gestures. Moving the hands while signing can signify ongoing action, repetition, and other grammatical structures. Furthermore 'non-manual signals' such as facial expressions can add meaning. For example, leaning forward while raising the eyebrows can turn a statement into a question. These secondary gestures convey information that can dramatically alter the meaning of signs and progress has been made in detecting non-manual gestures in sign language. Erden and Sclaroff (2002) use a three-dimensional head tracker (La Cascia et al. 2000) to detect head shakes in sign language. These methods do not, however, combine the manual (hand signs) and non-manual (head and face gestures) in the analysis.

Hand and head gestures have also been used in a variety of HCI applications. Hand-based interfaces include pointing to select virtual objects (Colombo et al., 2003), gestural interfaces to computer games (Kang et al. 2004), and using the hand as a three-dimensional mouse (Nesi and Del Bimbo 1996). Examples of head gestures in interfaces include Davis and Vaks (2005) who present a user interface for a responsive dialog-box agent that uses real-time computer vision to recognise user acknowledgements from head gestures and where a nod means "yes" and a shake means "no". El Kaliouby and Robinson (2003) describe a similar affective message box, which employs a real time gesture recognition system as its input modality, while Deniz et al. (2004) consider gestures as a method for human-robot interaction.

Sign language and HCI systems typically assume the user is communicating directly with the device, providing a clear, full-face image and allowing facial features to be used to support tracking and recognition. As a result, such systems tend to be two-dimensional, operating on features of the image rather than of the viewed world. Kawato and Ohya (2001) propose an approach for detecting nods and shakes in real time from a single colour video stream, which depends on detecting and tracking a point between the eyes. Kapoor and Picard (2001) describe a system that detects head nods and head shakes in real-time using an infrared sensitive camera equipped with two concentric rings of infrared LEDs to track participants' pupils, and eye tracking is also employed in the system proposed by Tang and Rong (2003). Morimoto et al (1998) employ an explicit three-dimensional model, describing the participant's face as a planar surface and basing the recognition of head gestures on changes in the parameters of that plane. The plane representation is only a very crude approximation to the human face and captures only a small part of the facial variation that takes place during conversation. Moreover, the face must be (almost) entirely visible if a plane is to be fit with the necessary degree of accuracy.

In order to construct a vision system that is capable of assisting in the analysis of video corpora of natural conversation there are a number of issues that need to be addressed. The main issues arising are how to:

- Provide the ability to rapidly search large quantities of video for potential gestures of interest.
- Analyse these gestures to extract meaningful descriptors of the motion involved.
- Identify common classes of gesture, on the basis of these descriptors.
- Combine the gestural information with the other modalities in the corpus, particularly the spoken language.

These issues may be mapped directly to the flow of information through a computer vision system, illustrated in Figure 1. Input is a video sequence (or set of sequences from several view points). The first process is to extract gestures from the input video to give short sequences containing a single gesture. These sub-sequences are then processed further to extract features that describe the gesture. For example, a head nod might be described by its duration, its amplitude (amount of motion), and its frequency (speed). A video corpus will provide a large number of instances of gestures, and these can then be analysed further on the basis of the extracted features. Of particular interest here is the classification of gestures into categories or groups which can be examined by expert linguists to determine their linguistic relevance in the context of conversation.

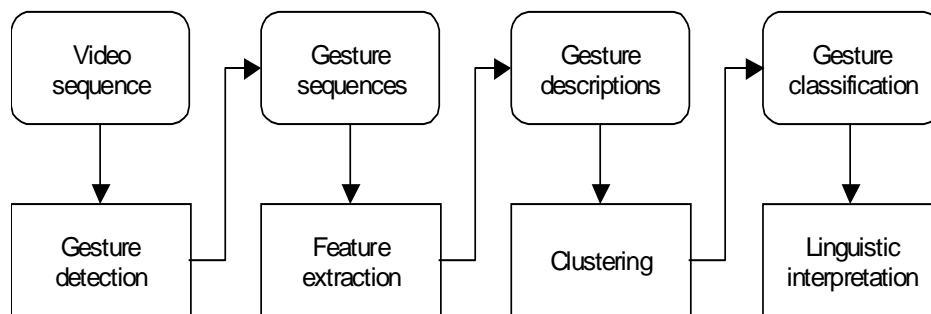


Figure 1. System overview showing the data (top) and processes (bottom) in the automated analysis of conversational gesture.

Gesture Detection

The first task is the detection of gestures from a video sequence. This may be done manually, but is time consuming and so not feasible for large corpora. Gesture detection and feature extraction need not be mutually exclusive – methods which describe gestures can be used to find them by looking for sections of video where the description of the person’s motion matches the characteristics of the gesture. However there are some methods which give much less detailed information which might be applied to the detection of gesture. This is advantageous since these methods are often quite simple, and therefore can efficiently be applied to large corpora. There are therefore, a range of possibilities, from applying detailed (and expensive) analysis to the entire video corpus, to applying a very coarse (but efficient) process to identify sequences of interest for more detailed analysis. At one extreme this can become prohibitively expensive in terms of computation, while at the other there may be insufficient information to reliably detect and describe gestures.

As an example of a method that may be used to efficiently identify gestures, we consider the use of wavelets to locate head nods in a video sequence, one frame of which is shown in Figure 2. The video frames are first decomposed using a discrete wavelet transform (Pittner and Kamarth 1999). This gives four wavelet components, and the second and third components represent vertical and horizontal frequency respectively. Statistical moments (Nixon and Aguado 2002) are then computed for each of these wavelet components. Figure 3 shows the first three moments of the vertical wavelet component for a sequence containing two head nods. The two nods are clearly seen as step changes in the moment values.

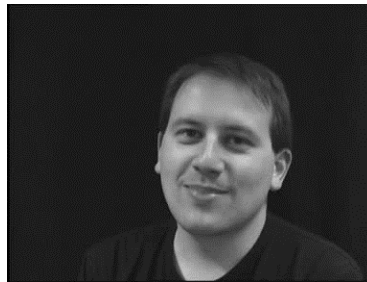


Figure 2. One frame from an example video sequence.

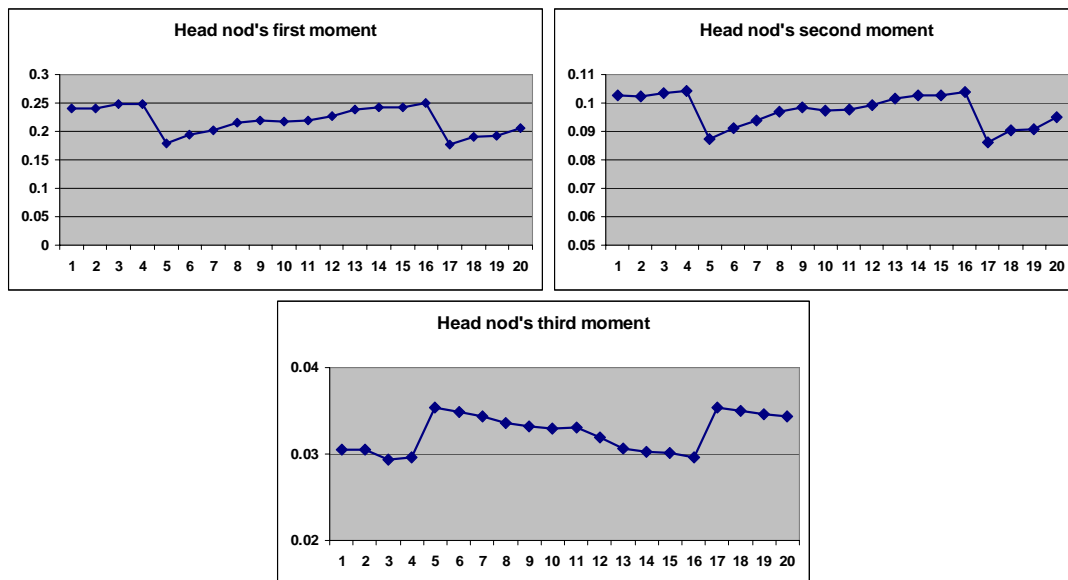


Figure 3. The first three moments of the vertical wavelet component.

The discrete wavelet transform is capable of rapidly extracting accurate estimates of motion across the input image (or image region) in two key directions. The data presented above strongly suggest that these motion estimates will support automatic detection of head nods and shakes. The wavelet method detects motion, and cannot discriminate between motion of a person's head and motion of other objects and the scene. It can, however, rapidly detect candidate sections of video for further, more detailed, analysis.

Feature Extraction

In order to accurately describe and analyse gestures, we model the head and use this model to track the face through a video sequence. Figure 4 shows an example of a three-dimensional model tracking a person's face. This model has six parameters that control the location and orientation of the model in space. Further parameters (15 in this example) control the changes

in appearance and shape of the model. This model is built from training data using principle component analysis, in a similar way to the two-dimensional active appearance models of Cootes et al. (2001).



Figure 4. Two frames from a video sequence with a three-dimensional model fitted to them.

By analysing the orientation parameters, which represent rotations of the head, we can characterise head nods and shakes. The results of this are shown in Figure 5. This figure shows the pitch (rotation about an axis roughly ‘through the ears’) and yaw (rotation about an axis vertically through the head) of a short sequence containing a nod (frames 20-60), followed by a head shake (frames 70-110), and then a smaller nod (frames 125-135). These gestures can be seen as oscillations in the relevant angles.

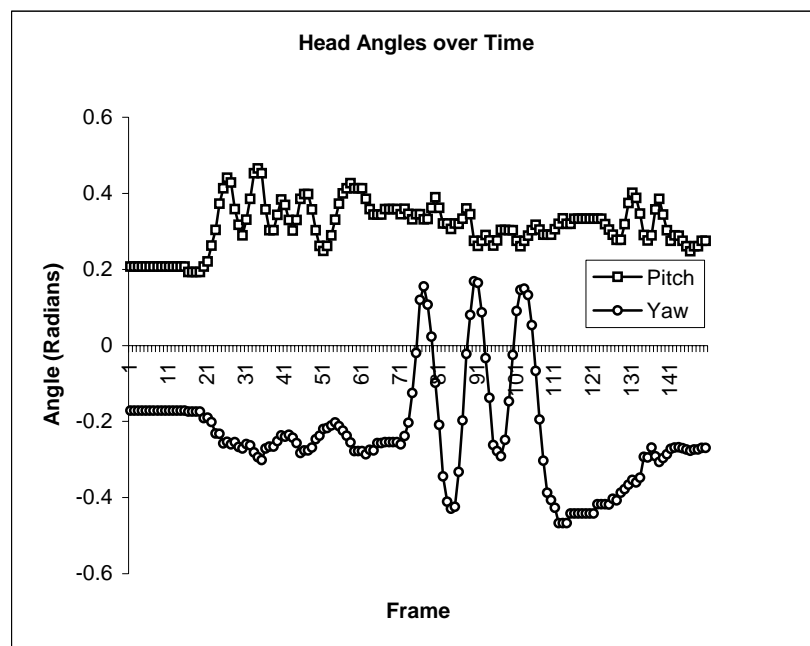


Figure 5. Head angles over time, with a nod, followed by a large shake, then a smaller nod

Clustering and Linguistic Interpretation

The final stage of processing is to relate the gestures, each described by a small number of features, to linguistic categories. The methodology we have chosen combines automated clustering methods with linguistic classification of the gestures. The three general approaches that might be used are to:

1. Apply automated clustering techniques, and then to analyse the clusters produced to determine their linguistic significance (if any).
2. Group the gestures based on a classification from expert linguists, and then try to learn the features that may be used to distinguish between the groups.
3. Apply automated and expert clustering in an iterative manner, with each informing and refining the other.

The first two options are unsuitable since they give one approach dominance over the other. In the first instance the linguistic analysis is constrained by the results of the automated processing, while the second case requires *a priori* knowledge of the number, nature, and contents of the final clusters. With an iterative approach, an initial estimate of the clusters may be made, and then refined by expert analysis. This can then inform development of the clustering techniques, and suggest new features for the analysis of gesture.

There are a large number of techniques available for automated clustering. Perhaps the most common is *k*-means where a number of clusters (*k*) is chosen, and then *k* points from the sample set are picked (at random or through structured sampling) to represent the cluster centres. The remaining points are then allocated to the cluster centre nearest them. This gives an initial classification and the centroid of each cluster becomes its new centre estimate. This process of cluster assignment then centroid estimation is repeated until the clusters are stable. More advanced methods, such as expectation-maximisation approaches (Dempster et al., 1977) determine the number of clusters as well as their contents, and so do not require that the value of *k* be known in advance.

To examine the feasibility of this methodology we have started with a simple feature to describe head nods – the duration of the nod. Duration is the simplest feature to extract from a suitably sized dataset of gestures. An initial clustering was made using *k*-means, with *k* = 4. The clusters were not individually meaningful, but it was noted that the two clusters that related to a longer head-nod duration contained a greater proportion of non-verbal gestures. As a result a revised classification was made, with *k* = 2. The results are shown in Table 1. A chi-square analysis of this data gives $\chi^2 = 2.62$ which is close to the 10% level of confidence critical value of 2.71. This indicates that there may be some relationship between duration and verbalisation – gestures associated with a verbal response may be shorter than non-verbal gestures. The data, however, is not strong enough to draw statistically significant conclusions at the 90% confidence level.

	Short	Long
Verbal	122	66
Non-verbal	47	39

Table 1. Number of gestures classified as short or long, and as verbal or non-verbal.

Within the verbal gestures it is possible to distinguish between gestures that act as a backchannel and those that do not. This information is shown in Table 2. Here, chi-square analysis gives $\chi^2 = 10.57$, which is significant at a 99% confidence level (a critical value of 6.63). This indicates that there is a statistically significant relationship between the duration-

based clusters and the role of gesture in conversation, at least for verbal gestures. On the whole, gestures that act as a backchannel tend to be shorter than those that do not.

	Short	Long
Backchannel	113	50
Non-backchannel	9	16

Table 2. Verbal gestures classified as short or long, and as backchannel or not.

Conclusion

The construction and analysis of multi-modal linguistic corpora is an important and open research challenge. Tools are required which allow social scientists to create and annotate such corpora. As much of the raw data is visual in nature these tools must incorporate techniques developed in computer vision. Gesture recognition has been well-studied, but not in this domain. As a result, no directly applicable system exists, but many valuable components are available.

We need to select and combine components to create a head nod/gesture system capable of providing results that tally with the linguistic interpretations made by corpus linguists. A number of methodologies present themselves. We believe the most effective route is to begin an iterative programme of linguistic and feature-based classification and refinement, both of the classes identified and the linguistic and visual features used to separate them. We have demonstrated range of techniques available and presented initial results in this direction.

References

- Bennett, M. & Jarvis, J. (2001). The communicative function of minimal responses in everyday conversation. *The journal of social psychology* 131, 4: 519-523.
- Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus Linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.
- Brown, R. (1986) *Social Psychology* (2nd ed.), New York: Free Press.
- Carter, R. & McCarthy, M. (2006). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.
- Colombo, C., Del Bimbo, A., and Valli, A. (2003) "Visual capture and understanding of hand pointing actions in a 3-D environment", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 33 (4), pp. 677-686.
- Cootes, T., Edwards, G. and Taylor, C. (2001) "Active appearance models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23 (6), pp. 681-685.
- Davis, J. and Vaks, S. (2001) "A perceptual user interface for recognizing head gesture acknowledgements", *Proceedings of the 2001 ACM Workshop on Perceptive User Interfaces*, Orlando: <http://www.cse.ohio-state.edu/~jwdavis/publications.html>
- Dempster A., Laird N. and Rubin D. (1977) "maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, 39, Series B., pp 1-39.

- Deniz, O., Falcon, A., Mendez, J. and Castrillon, M. (2004) "Useful computer vision techniques for human-robot interaction", *Proceeding of the 1st International Conference on Image Analysis and Recognition*, Porto, Portugal: <http://mozart.dis.ulpgc.es/Gias/Publications/iciar04.pdf>
- Duncan, S. & Niederehe, G. (1974). On signalling that it's your turn to speak. *Journal of experimental social psychology* 23, 283-292.
- Efron, D. (1941). *Gesture and environment*. New York: Kinds Crown.
- Ekman, P. and Friesen, W. (1969) "The repertoire of nonverbal behaviour: categories, origins, usage and coding", *Semiotica*, vol. 1, pp. 49-98.
- El Kaliouby, R. and Robinson, P. (2003) "Real-time head gesture recognition in affective interfaces", *Proceedings of the 2003 IFIP TC13 International Conference on Human-Computer Interaction*, pp. 950-953, Zurich: IOS Press.
- Erdem, U. and Sclaroff, S. (2002) "Automatic detection of relevant head gestures in American sign language communication", *Proceedings of the 6th International Conference on Pattern Recognition*, pp. 460-464, Quebec: IEEE Computer Society.
- Fellego, A.M. (1995). Patterns and functions of minimal response. *American Speech* 70, 186-199.
- Gardner, R. (1997) "The listener and minimal responses in conversational interaction", *Prospect*, vol. 12 (2), pp. 12-30.
- Gardner, R. (1998) "Between speaking and listening: the vocalisation of understandings", *Applied Linguistics*, vol. 19, pp. 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*, vol. 3 (11), pp. 419-429.
- Haiman, J. (1998) "The metalinguistics of ordinary language", *Evolution of Communication*, vol. 2 (1), pp. 117-135.
- Kang, H., Lee, C.W. and Jung, K. (2004) "Recognition-based gesture spotting in video games", *Pattern Recognition Letters*, vol. 25 (15), pp. 1701-1714.
- Kapoor, A. and Picard, R. (2001) "A real-time head nod and shake detector", *Proceedings of the 2001 ACM Workshop on Perceptive User Interfaces*, Orlando: <http://vismod.media.mit.edu/tech-reports/TR-544.pdf>
- Kendon, A. (1972) "Some relationships between body motion and speech", *Studies in Dyadic Communication* (eds. Siegman, A. and Pope, B.), pp. 177-210, Elmsford, New York: Pergamon.
- Kendon, A. (1980) "Gesticulation and speech: two aspects of the process of utterance", *The Relationship of Verbal and Non-verbal Communication* (ed. Key, M.), pp. 207-227, The Hague: Mouton.
- Kawato, S. and Ohya, J. (2000) "Real-time detection of nodding and head-shaking by directly detecting and tracking the 'between-eyes'", *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 40-45, Grenoble, France: IEEE Computer Society.
- La Cascia, M., Sclaroff, S. and Athitsos, V. (1999) "Fast, reliable head tracking under variable illumination", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21 (6), pp. 322-326.
- Maynard, S. (1989) *Japanese Conversation: Self-contextualization through Structure and Interactional Management*, Norwood, New Jersey: Ablex.
- Maynard, S. (1990) "Conversation management in contrast: listener response in Japanese and American English", *Journal of Pragmatics*, vol. 14, pp. 397-412.

- Maynard, S. (1997) "Analyzing interactional management in native/non-native English conversation: a case of listener response", *International Review of Applied Linguistics*, vol. 35, pp. 37-60.
- McCarthy, M. (2001) *Issues in Applied Linguistics*, Cambridge: Cambridge University Press.
- McClave, E. (2000) "Linguistic functions of head movements in the context of speech", *Journal of Pragmatics*, vol. 32 (7), pp. 855-878.
- McEnery, T. and Wilson, A. (1996) *Corpus Linguistics*, Edinburgh: Edinburgh University Press.
- Morimoto, C., Koons D., Amir A. and Flickner M. (1998) "Real-time detection of eyes and faces", Proc. Workshop on Perceptual User Interfaces pp. 117-120.
- Mott, H. and Petrie, H. (1995) "Workplace interactions: women's linguistic behaviour", *Journal of Social Psychology*, vol. 14, pp. 324-336.
- Nesi, P. and Del Bimbo, A. (1996) "A vision-based 3-D mouse", *International Journal of Human-Computer Studies*, vol. 44 (1), pp. 73-91.
- Nixon, M. and Aguado, A. (2002) *Feature Extraction and Image Processing*, Oxford: Elsevier.
- O'Keeffe, A. & Adolphs, S. (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.
- Ong, S. and Ranganath, S. (2005) "Automatic sign language analysis: a survey and the future beyond lexical meaning", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27 (6), pp. 873-891.
- Pittner, S. and Kamarth, S. (1999) "Feature extraction from wavelet coefficients for pattern recognition tasks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21 (1), pp. 83-88.
- Roger, D. and Neshover, W. (1987) "Individual differences in dyadic conversation strategies: a further study", *British journal of Social Psychology*, vol. 26, pp. 247-255.
- Schegloff, E. (1982) "Discourse as interactional achievement: some uses of "uh huh" and other things that come between sentences", *Analyzing Discourse, Text, and Talk* (ed. Tannen, D.), pp. 71-93, Washington DC: Georgetown University Press.
- Tang, W. and Rong, G. (2003) "A real-time head nod and shake detector using HMMs", *Expert Systems with Applications*, vol. 25, pp. 461-466.
- Tottie, G. (1991). Conversational style in British and American English: The case of backchannels. In Aijmer, K. & Altenberg, B. (Eds.), *English corpus linguistics*. London: Longman. 254-271.
- Wilcox, S. Language from gesture. *Behavioural and Brain Sciences* 24: 4, 525-526.
- Yngve, V. (1970) "On getting a word in edgewise", Papers from the 6th Regional Meeting of the Chicago Linguistic Society. Chicago: Chicago Linguistic Society.

ⁱ This notion of backchannels has been supported by an extensive body of research, although various alternative terms have been used to describe them, including 'accompaniment signals' (Kendon, 1967), 'assent terms' (Schegloff, 1972), 'listener responses' (Roger, Bull and Smyth, 1988) and 'minimal responses' (Fellegly, 1995), (all reported in Bennett & Jarvis, 2001).