In: Lüdeling, Anke / Kytö. Merja (eds) Corpus  Linguistics. An International Handbook. Mouton de Gruyter. Berlin: 207-225

# Multimodal Corpora

Jens Allwood

## Introduction

The structure of this paper is the following. In section 1, multimodal corpora are defined and described, in section 2, reasons are given for why multimodal corpora are created, and in section 3, there is a discussion of some issues to keep in mind when creating and analyzing a multimodal corpus. There is also a discussion of some research directions. In section 4, possible applications are mentioned and section 5, finally, contains some concluding remarks.

## 1. What are multimodal corpora?

The Latin word ”corpus” (body) is used to metaphorically describe a collection of language and communication data, see Lewis (1966). In what follows, I will assume that the corpus is stored on a computer, i.e. a digitized corpus. A digitized corpus is, thus, a kind of database of language related material. Although computer based corpora were planned in the 1940’s (cf. http://en.wikipedia.org/wiki/Roberto_Busa), the first appeared in the 1950´s and contained written language excerpts. They were an attempt to replace earlier manually collected and stored sets of written excerpts with a set stored on a computer. During the 1960’s also corpora of transcriptions of spoken language appeared. For a history of corpora, see McEnery and Wilson (2001) and the articles in section I. If we want to find out when the first multimodal corpora appeared, the answer to this question is dependent on how we define a “multimodal corpus”. In the widest sense, it may be a collection of analog films, which are registered in a paper file or on a computer**.** In a slightly more narrow sense, which will be the sense I will discuss here, it would only include material, that has been digitized, i.e., the films would have to be digitized rather than just available in an archive. A first attempt at a definition might now be to say that a multimodal digitized corpus is a computer-based collection of language and communication-related material drawing on more than one sensory modality or on more than one production modality (see below). Depending on how narrow we want our sense of “corpus” to be, we might then, for example, revise this definition to say that a multimodal corpus is a digitized collection of language and communication-related material, drawing on more than one modality. In a more narrow sense, we might require that the audiovisual material should be accompanied by transcriptions and annotations or codings based on the material. This definition is more narrow, since there is a specification of the nature of the language and communication related material, i.e. it should contain recordings, transcriptions and annotations. The first definition leaves the nature of the corpus open.

Examples of multimodal corpora might, thus, be a digitized collection of texts illustrated with pictures and/or diagrams or a digitized collection of films with associated transcriptions of the talk in the films.

Another issue, already hinted at, is what is meant by multimodality. The term ”modality” can be used in many ways, but the definition we will adopt is that “multimodal information” is information pertaining to more than one “sensory modality” (i.e., sight, hearing, touch, smell or taste) or to more than one “production modality” (i.e., gesture (the term “gesture” will, in this paper, be used in the sense of any body movement), speech (sound), touch, smell or taste). If we assume that there are five or more sensory modalities (e.g. vision, hearing, touch, smell, and taste) only two of these have really been made use of so far in multimodal corpora, namely vision and hearing (which corresponds to the production modalities of gesture and speech in face-to-face communication). The term “multimodal” can, thus, be contrasted with the term “multimedial” which has a slightly different sense, relying on the notion of “medium”. This term can also be used in many ways and is sometimes taken in more or less the same sense we have given multimodal above. However, in order to maintain a contrast, we will define a “communication medium” as the physical carrier of multimodal information. Thus, the medium for sight is light waves, the medium for hearing is sound waves, the medium for touch is physical pressure and the medium for smell and taste various types of chemical molecules.

Multimodal corpora are multidimensional, not only from a modality point of view, but also from a semiotic point of view. Often, all the three Peircean information carrying relations are present, i.e. index, icon and symbol (cf. Peirce, 1955). Even though still pictures and motion pictures are themselves iconic in nature, both types can iconically represent indexical, symbolic and even iconic information (a picture of a picture). This is also true of sound recordings, where the recorded sound is an iconic representation of the original sound which in itself can contain indexical, symbolic and even iconic information (e.g. a sound recording of an imitated sound).

Besides what is iconically represented in this way by audio- or video recordings, the corpus can contain symbolic textual information which can add to, supplement and complement the recordings. Let us distinguish three cases:

(i) Texts describing pictures.

Pictures give concrete iconic details, e.g. a particular brown horse on a particular field, while words give more abstract symbolic information. The word “horse” does not tell us what color the horse has, but it is impossible to depict a horse without depicting a particular color. Words can add focus, identification and perspective to a pictorial representation. In fact, most existing multimodal corpora rely on textual identifying information in searching the corpus. This is so, since present technology mostly does not really allow efficient search using the iconic elements themselves. So texts in a multimodal corpus can be used for identification but also to give historical or background information.

(ii) Audio-video recordings with annotations or codings.

A special case of texts describing pictures is provided by text (usually called annotations or codings) which gives a kind of descriptive running commentary on what occurs in the recording. This kind of textual information is often used to describe gestures, prosody or other aspects of sound quality. It can also be used to capture features of context or various types of semantic-pragmatic information.

(iii) Audio-video recordings with transcriptions.

A second special case (in fact a special case of annotations and codings) is provided by so called transcriptions, i.e. text which gives a more direct representation (usually symbolic) of what is said or done. The most common type of transcriptions (cf. Allwood et al., 2000) are related to audio recordings and represent segmental speech sounds, leaving out gestures, prosody and other aspects of sound quality. There are, however, special types of transcription which include such information. For gestures, see, for example, Birwhistell (1952) or Laban

(1974) and for prosody, see, for example, Svartvik & Quirk (1980) or Brazil (1985). The difference between transcriptions and annotations (or codings) lies in the attempt of transcription to give a direct moment to moment representation of what is said, rather than to give a more indirect and mostly less continuous description of certain properties of what is said or done.

For all these three types of textual information, the question how they relate to the recorded material, thus, may be raised. To ease comprehension, the general principle adopted for all three types is that of spatio-temporal contiguity. A text occurs at the same point in time as the event it describes or represents. Sometimes, it is even placed at the same point in space, as for example when a label for an object or the name of a person is placed on or in the immediate vicinity of the iconic object it identifies. Usually, however, only temporal contiguity is maintained. When temporal contiguity concerns the relation between transcribed speech (or gesture) and recorded speech (or gesture), it is often referred to as "synchronized alignment" of recording and transcription. The degree of synchronization can vary from the subtitles conventionally used in translating commercial movies which usually occur on an utterance level, in such a way that the whole transcribed utterance is visible as it is being said, to the more fine-grained synchronized alignment used in studies of phonetics, where each phoneme is aligned with a feature in an indexical representation of the acoustic features of the utterance. What synchronization means is that for every part of the transcription (given a particular granularity), it is possible to hear and view the part of the interaction it is based on and that for every part of the interaction, it is possible to see the transcription of that part.

In general, synchronization of information in different modalities has turned out to be a difficult problem in assembling a multimodal corpus. It concerns not only the relation between text and iconic representation but also between different means of recording in the same or different modalities. For example, how should several cameras recording the same event from different perspectives be synchronized or how should several microphones recording a multiparty conversation be synchronized and how should sound and pictures be synchronized? There are two main ways of handling the problem:

(i) Synchronization can be done by a computer program, already while making the recording. This is the most convenient solution (cf. the AMI project, http://www.amiproject.org, and The CHIL project, http://chil.server.de/servlet/is/101/, or Zhang et al., 2006, for interesting examples of how this can be done).

(ii) Synchronization can also be done after the recordings have been made.

In addition, it is possible to mix these two approaches, so that some synchronization is done during the recording and some more is done using the finished recordings.

The kind of multimodal corpus we will be mostly interested in below can be characterized as a digitized collection of audio- and video-recorded instances of human communication connected with transcriptions of the talk and/or gestures in the recordings. The two modalities are, thus, hearing and vision. There are audio recordings of the speech and there are video recordings of the body movements of the participants in the interaction. In addition, there are the transcriptions, which, as we have mentioned, are a kind of visual symbolic representation of the speech (and more rarely of the gestures that occur in the recordings). The form of connection between the transcriptions and the material in the recordings can vary from just being a pairing of a digitized transcription with a digitized video or audio recording (both recording and transcription exist but they have not yet been synchronized) to being a complete temporal synchronization of recordings and transcription.

## 2. Why multimodal corpora?

The basic reason for collecting multimodal corpora is that they provide material for more complete studies of “interactive face-to-face sharing and construction of meaning and understanding” which is what language and communication are all about. Such studies are not fully possible in corpora which contain linguistic material of a less comprehensive kind, since much of the sharing and construction of information is done multimodally through a combination of gestures and speech (including prosody), i.e. concern processes which integrate multimodal information (in perception and understanding) and distribute information multimodally in production and are processes of which we often have a low degree of awareness.

Examples of such, often very automatic types of processes can, for example, be found in the head nods by which one speaker gives feedback to another speaker (while he/she is speaking), or in the head movements which a speaker uses to elicit attention from other interlocutors. For an account of processes of this kind, see Goodwin (1981) and Allwood (2001a, 2002).

Another reason is that speech and gestures, unlike written language, are transient objects: they disappear when they have been produced. Given that our intuitions about the nature of naturalistic speech and gesture are usually very unreliable, there is a need for a less transient type of object to study. A corpus of multimodal communication is this kind of object.

Simplifying the matter slightly, the study of multimodal communication enables us to highlight how we continuously in interaction incrementally combine communicative actions with other instrumental actions in order to share information, e.g. compare a situation where I pour coffee into a cup and hand it over to you saying *coffee* with a rising intonation with a situation where I lift an empty cup directing it to you (you are sitting next to the coffee pot) and say *coffee* with a falling intonation. The former might be construed as an offer while the latter perhaps might be construed as a slightly impolite demand. In both cases, our interpretation is dependent on an integration of signaled verbal linguistic information with indicated and displayed non-linguistic actions (cf. Allwood 2002).

The general problem of how information from communicative actions is combined with information from instrumental actions which are not primarily communicative, has not been extensively studied so far. However, the more limited problems of describing the function of visual communicative gestures (cf. Poggi 2002, Kendon 2004, Argyle 1988, Allwood 2001a) and describing the integration of words, with prosody and gestures have been somewhat more studied (cf. Allwood 2002).

If we turn to the functions of the different modalities in communication, we may say that vocal verbal elements (what is usually captured in conventional writing systems) are our primary source of factual information. In addition, this vocal verbal information is often supplemented by conventional symbolic gestures and illustrative iconic gestures. Vocal verbal information is also used for communication management, prosody gives us information about information structure and emotions/attitudes, while gestures primarily give us information both about emotions/attitudes and communication management. Sometimes the word “social” is used for aspects of communication that do not relate to factual information. Even though this is not a very good terminology (since communication is always social and what is social often encompasses factual information and what is non-factual is not always social), it should be clear that the “social” functions of language, to a very high degree, rely on information which is gestural and prosodic, i.e. necessitate a multimodal approach to communication.

In any case, we can see that, given the central role of prosody and gestures for the communication of emotion and attitude, studies of affective behavior and affective display, which are to have ecological validity, are highly dependent on creation of reliable naturalistic multimodal corpora.

In the following example we illustrate some of the types of behavior and processes that are involved in multimodal communication.

Example 1. Video-based analysis of vocal verbal and gestural elements in a case of hesitation, i.e. OCM “own communication management”

Speaker: *å där så de e som* *<u>en e // sportspår</u>* *där som vi springer*

*and there so it is like* *<u>a eh // sportstrack</u>* *where we run*

A closer look at the relationship between vocal words and gestures in the OCM part of the utterance the phrase (*en e // sportspår)* is provided in table 1.

| **Speech** | en | e | // | sportspår |
|---|---|---|---|---|
| **Type** | Indef article | OCM word | pause | Noun |
| **Gesture** | hand circling, illustrating track | turns away head and gaze | | Head and gaze back |

Table 1: Multitrack annotation of an example of own communication management

If we start by examining the temporal relation between the vocal-verbal and gestural production, we see that an illustrating gesture occurs before the OCM word *e* and pause, which both precede the word *sportspår (sportstrack)*. We interpret this as indicating that the speaker has a problem in choosing and producing the right word and that this is reflected in the use of the OCM word and pause to gain time. We can also see that the gesture occurs as the article preceding the OCM word is produced. The gesture might in this case be an illustrating iconic gesture, which could have occurred even if the speaker had no need for support in finding the word, but it might also have a self-activating word finding function for the speaker. The gesture also serves to keep the floor and to give a clue about the meaning of the coming word to the listeners.

Simultaneously with the OCM word and pause, the speaker turns his head and gaze away from the interlocutors, indicating memory search and giving further support for turnkeeping. When he produces the noun he moves his head back facing the interlocutor, indicating that the memory search is completed.

Thus, the example indicates that normal face-to-face communication contains a wealth of multimodal information, vocal-verbal as well as gestural and that the temporal relation between the modalities is not simple. The example is fairly typical of the complex relation between speech and gesture, the study of which is facilitated by the kind of data that are available in a multimodal corpus.

A multimodal corpus, in this way, gives one an opportunity of capturing not only written language (or a written transcription of spoken language) but provides an opportunity to include information of a contextual and cultural kind. This means that multimodal corpora are excellent instruments for a more holistic documentation of cultural and linguistic processes (as is currently going on in relation to the endangered languages of the world, see article 23). It also means that theories of language and communication, by giving access to more relevant data, potentially can provide a better and more correct description, understanding and explanation of the nature of language.

In line with what has been said above, another area that currently is driving the need for multimodal corpora is the construction of so called embodied conversational agents or avatars, i.e. artificial computer-based communicators that have a more or less artificial face and body. Such avatars are today becoming more and more human-like. This means that they are being equipped with the same kind of communicative behavior and communicative functions that humans have. Among other things, this means that they are capable of showing emotions and attitudes. For interesting examples of avatars (or ECAs) of this type see the HUMAINE network (http://emotion-research.net), Cassell (2000), Gratch et al. (2006).

Since the approach relies on simulating multimodal human communication, this means that there is a great need for more exact information of this type. The primary source of this information is corpora of multimodal communication, cf. Martin et al. (2005).

Also in the area of computer mediated communication (CMC) use is made of corpora of multimodal communication. An important goal of CMC is to facilitate human communication in various ways, e.g. in order to bridge gaps of space and time or to provide summarization of meeting contents. (Cf. the AMI project at http://amiproject.org for interesting examples of techniques to create so called "virtual meeting rooms", Nijholt et al. (2006a), and article 21). If this is to be done efficiently, the suggestions made have to be based on studies of actual human communication. Otherwise, there is an obvious risk that what will be produced will never be used. Again, the key is to have information on actual multimodal human communication available in corpus form.

## 3. Creating a multimodal corpus

Let us now take a look at some of the considerations that need to be kept in mind, in constructing a multimodal corpus based on recordings of face-to-face communication. We will consider the following issues (some of which are also covered in article 33):

1. What should we record?
2. How should we record?
3. How should we keep track of the recordings?
4. Should we transcribe and, if so, how?
5. How should we keep track of the transcriptions?
6. How should we analyze recordings and transcriptions?
7. What should we analyze?

### *3.1 What should we record?*

What we should record depends on the purpose of our investigation (cf. also article 11). There are many possible criteria for deciding on what data should be recorded and included in the corpus. What sampling criteria for the corpus (e.g. speaker characteristics like; age, sex, social class, personality type, level of education, ethnic background, occupation or regional background) should be chosen is dependent on what we are investigating and what is important for this investigation. However, since a corpus is often collected in order to be a resource for more than one purpose, many researchers are interested in getting a balanced sample in the sense of taking into account as many of the sampling criteria as possible. What this could mean is that our corpus should not only contain women but both men and women, not only children but children, adults and persons of old age, etc. In some cases, however, one does not want a balanced sample but rather a specialized sample of only women or only young men of working class background, etc.

Another problem that arises is the choice of persons within each category. For example, could we merely record the persons we happen to run into in each category? This is sometimes done and can provide interesting results. Another possibility is to use some type of random sampling within each category. This requires getting a list of the possible candidates and then using some method of random sampling to pick out the persons who will be recorded. A third possibility is to use some sort of strategic sampling, where the theory to be investigated decides what data to choose.

Instead of basing our corpus on speaker characteristics, we could choose to sample on the basis of social activity or type of organization. We could, for example, try to record activities related to, e.g., research, education or manual work (e.g. fishing, hunting, farming, crafts), industrial work, commerce, religious practice, healthcare, judicial (law) practice, entertainment, mass media, transportation, building, professional food, military or everyday life (including relaxation). Even if the main purpose in gathering data according to social activity would be to record communication in the mentioned activities irrespective of the personal characteristics of the participants, it may still be desirable to keep track of these characteristics. They will, however, usually be of secondary interest and we would normally accept an activity-based corpus that had more women than men or more middle class people than upper class people, etc., since we are primarily interested in the nature of the activity-based interaction rather than in the influence of the participants' personal characteristics on the interaction.

Unfortunately, the answer to the question of what a balanced sample is is even more unclear in relation to social activities, than it is in relation to personal characteristics. The list of activities given in the previous paragraph represents an effort to find such a list but there is no guarantee that something important is not missing. In many cases, the nature of the list is also culture-dependent and is therefore going to change with time and be different in different cultures.

### *3.2 How should we record?*

Once we have decided what to record, we have to decide how to do this. Since this topic will be covered in more detail elsewhere in this book (see articles 32 and 33), I will restrict myself to a few of the relevant issues.

The first issue concerns choice of medium for registering data. Should we, for example, rely on memory, use written notes (the classical anthropological field notes) or use audio and video recordings? Probably, we will end up with a combination of all these three methods.

When it comes to recording, we are faced with many options and questions that need to be answered regarding choice of equipment. For example regarding

- choice of audio/video recorder
  analog/digital
  size of recorder
  number of channels
- choice of microphone (directed, wide angle, radio)
- choice of camera
- choice of lens
- choice of lighting
- choice of view, focus, panorama, etc.
- choice of tape/digital memory

It is not always easy to make the right decisions since the relevant technology is changing very quickly. What is impossible today might very well be possible tomorrow.

Let me just stress a few points.

(i) It is very important to establish routines to be followed in making the recordings, e.g. concerning what to do before in preparation, what to do during the recording and what to do afterwards regarding storage and access.

(ii) In making video recordings that have as a purpose to document regularities and interactive mechanisms in normal communication, it is not a good idea to shift focus or to move the camera from one speaker to another. In order to capture the interaction, the camera should have a constant wide-angle view, trying to get a picture of all participants. If the focus keeps shifting from speaker to speaker, the interaction is lost. A compromise, if your resources allow you, is to have a second camera following the current speaker while keeping the first camera constant on the interaction. In editing the recordings, the second camera recording can then be synchronized with the first camera recording and inserted in a corner of the wide-angle view, thus providing both interaction and focus on speaker.

(iii) Another point concerning video recordings and the study of gestures is that we should be aware that many positions, e.g. sitting down around a table, limit both our freedom to gesture and our ability to record and observe what gestures occur. As usual, the purpose of our investigation is important here. If we only want to study facial gestures and hand movements, probably the rest of the body is less important. But if we want a more holistic impression of multimodal communication, perhaps the best choice is to record communicators who are standing up and moving about.

(iv) In general, as recording of multimodal communication is getting more sophisticated, we are presently moving from using one microphone or camera to the use of several microphones and cameras. As already discussed above, this raises the problem of synchronization of recording devices while the recordings are being made. This area is currently under quick development and many new computer-based algorithms for synchronization of data are currently being tried out, cf. the AMI project (http://www.amiproject.org). It also raises the problem of how the data from several recordings is going to be presented to a human researcher. There is a need for more research on how to best visualize complex multimodal data in such a way that the constraints and capacities of human cognition are respected (cf. Nijholt et al. 2006b). This in one of the concerns motivating the suggestion made in point (ii) above.

(v) Finally, there is a never-ending tension between the desire to get high quality sound and video recordings (this is a must for many types of analysis) and the desire to get naturalistic recordings with high ecological validity (cf. Brunswick, 1969), outside of the studio, with no interference from the researcher. This tension is a challenge to try to optimize on both of these criteria as much as possible, i.e. always opt for as high quality as possible and as much ecological validity as possible.

### *3.3 How should we keep track of the recordings?*

All recordings that are made should be described with as much relevant background information as possible, e.g. date of recording, person who did the recording, type of recorder, length of recording, purpose of recording, what has been recorded, participants and characteristics of the participants (i.e., age, gender, social class. See section 3.1. above, concerning “personal characteristics”). This information is essential for the build-up of a systematic database containing the recordings. It is also very valuable to have for the person(s) who are going to transcribe the recordings, since they might not always be the same persons as the ones who made the recordings. The information makes it easier to understand what is going on, how many speakers there are, who is speaking, etc.

As recordings and transcriptions accumulate, the background information can be used to structure the actual physical archive or digital memory space allotted to the recordings (according to whether they are in analog or digital form). It can also be used for retrieval purposes, e.g. to find all recordings of auctions or of doctor-patient consultations. If personal characteristics rather than activities are the basis for the recordings, one might want to find all recordings of women who are middle-aged or men who are speakers of a particular dialect. What can be searched for and retrieved later when we want to use the corpus, entirely depends on what information about the recordings and transcriptions has been entered into the data base, when it was created.

If the recordings and transcriptions are not described, marked and registered (in the way suggested above) gradually as they come in, they usually end up in a more or less serious state of chaos, giving considerably more work to yet another person who gets the job of structuring the data and who, because of lack of information, usually is never able to do more than a partial job.

### *3.4 Should we transcribe and, if so, how?*

One of the first issues to decide on after we have made our recordings (or in some other way obtained audio/video material) is the question of whether the recordings should be transcribed or not? Transcription is certainly not the only way in which audio and video data can be studied. As we have seen above, one might, for example, attempt some sort of direct automatic computer-based analysis using speech recognition or attempt to annotate and code the material directly without also having made a transcription. For additional discussion, see article 32. There are many advantages of doing this, in terms of time and money, since transcriptions are both time-consuming and costly.

Another question concerns how the transcriptions should be produced, since transcriptions could, in principle, be produced online by speech recognition algorithms, as the recording is made. However, this is still not really possible, as there are too many mistakes in the output of the recognition algorithms. This means that transcriptions have to be made either in the traditional manual way by listening (observing) and transcribing what is heard or seen or by using automatically recognized speech (or gesture) as a first step and then correcting this step manually on the basis of listening and observation. The further synchronized alignment of transcriptions and recorded material can now take place, either by synchronizing the recording with the transcription (at the desired level of granularity) as the transcription is being made or by creating the synchronization at a later stage by matching transcription with recording. The second option is often a necessity if one wishes to align corpora of spoken language or multimodal communication, which were made before support of on-line synchronization was available. An interesting issue in this latter case is the creation of algorithms that can achieve synchronized alignment with recordings that have varying sound quality and contain overlapping speech.

We may here note that there probably are advantages of an automatic analysis in terms of not introducing biases, the most important of which probably derive from standard written language, involving such things as words space, spelling, capitalization and punctuation (none of which actually occur in normal face-to- face communication). An automatic type of analysis might enable us to get closer to the real properties of the acoustical and optical signal and, thus, perhaps also closer to more detail concerning how we perceive speech and gestures.

If we decide to transcribe, the question of what to transcribe needs to be answered. Should we attempt to transcribe the gestures (in the wide sense of all body movements), or should we restrict transcription to communicatively relevant body movements or perhaps even more to some more limited set of gestures, such as head or hand movements? Should we leave out gestures altogether and restrict transcription only to sound, perhaps limited to only some aspects of the speech signal?

If we decide to transcribe gestures (body movements), there are very many schemes available starting, for example, with Birdwhistell's proposals in the 1950's (cf. Birdwhistell, 1952). Laban's choreographically inspired proposals for posture and large scale gestures (Laban, 1974), the various proposals that exist for deaf sign language (Stokoe, 1978, Bellugi, 1972, Nelfelt, 1998, Prillwitz et al., 1989) or proposals inspired by the work of David McNeill (McNeill, 1979), Kita et al. (1997) and Månsson (2003).

Since a holistic transcription of gestures is so time-consuming, as to be almost impossible, most researchers usually decide on a system of annotation and coding that is oriented toward some communicative functions and thus does not cover everything (cf. for example, the gesture annotation schemas in Allwood, 2001b, and Allwood et al, 2005).

In a similar way, we have to decide how speech and sound are to be transcribed. Should we adopt a system which is close to standard orthography or should we adopt a system which captures as much fine-grained detail in the speech sounds as possible? The most detailed system of transcription is probably some version of the IPA (International Phonetic Alphabet). Related to the IPA (http://www2.arts.gla.ac.uk/IPA/ipa.html) there are then systems like SAMPA (www.phon.ucl.ac.uk/home/sampa/home.htm), **which make the IPA ASCII-compatible, and systems which extend the IPA and a host of less fine-grained systems than the IPA.** Among the less fine-grained systems, one might distinguish, e.g. phonemic systems and orthographic systems as well as several systems which propose a series of modifications of standard orthography, in order to get closer to various noticeable features of spoken interaction. Examples of two systems which employ such modifications of standard orthography are the system employed in Conversation Analysis (CA) (see Jefferson1984) and the GTS (Göteborg Transcription Standard) (see Nivre et al., 2004).

### *3.5.1 How should we keep track of the transcriptions?*

Most of what has been said above about keeping track of recordings is also relevant for keeping track of transcriptions. It is advisable either to initiate each transcription with a section giving all the relevant background information about the transcription or to have separate background posts or files for all transcriptions enabling you to retrieve all the transcriptions which have the prerequisite background information attached to them.

## *3.6 How should we analyze recordings and transcriptions?*

The most important factor in determining how we should analyze our recordings and transcriptions is, of course, our theoretical perspective and research objective, i.e., what we think we are investigating and why we are doing it. Since empirical reality potentially has a very large, probably infinite number of properties and relations, this means that any standard of transcription or schema of annotation/coding is dependent on a theoretical perspective and objective, which will select what empirical data are seen as relevant, determining what properties are picked up by the transcription or annotation and what properties are not picked up. This is so even if the adherents of the system of transcription or annotation claim they have no theory. Given the potential richness of empirical data, any transcription or coding must be the result of a selective procedure, even if the grounds of this procedure are not clearly articulated.

However, while fully acknowledging that the theoretical perspective, even if it is not very explicit, in the end is the most important determining factor, there are some general considerations that if taken into account will make a multimodal corpus more fully exploitable.

One such consideration is that the tools you employ for analysis should allow for simultaneous access to the relevant parts of the transcriptions and recordings. As we have already mentioned, a way to achieve this is to try to align the transcriptions and recordings so that there is direct relationship between what is said or done and what is transcribed, i.e., in reading the transcription on-line you are also able to open one or more other windows on your computer monitor, where you can see the relevant part of the video recording and hear the corresponding part of the audio recording. This means that there is synchronization between transcriptions and audio/video recordings. All three types of files are time stamped and the time points can be used for temporal alignment.

The synchronization between transcription and audio/video recordings can then further be extended so that all types of analysis that are done using the recordings or using the transcriptions are made accessible together with recordings or transcriptions. For example, we might be able to open windows for acoustic analysis, gesture analysis or functional analysis that are synchronized with the files containing recordings or transcriptions. With tools allowing this type of analysis, we can avail ourselves of the information available in the multimodal corpus, to a greater extent. For this reason, during the last few years, several tools for analysis of multimodal corpora have been developed which have some of the characteristica described above concerning synchronization between recordings, transcriptions and types of analysis. Some of these are:

ANVIL - http://www.dfki.de/~kipp/anvil/
MULTITOOL - www.ling.gu.se/projekt/tal/multitool

WAVESURFER - http://www.speech.kth.se/wavesurfer
NITE/MATE - http://nite.nis.sdu.dk/ and http://mate.nis.sdu.dk/

All systems have their strong points and their weak points and no system exists yet that has all desirable qualities. Most of the systems run better on PC than on Mac computers.

A second general consideration concerns whether the analysis should be done manually or automatically. If done automatically, it could, for example, be based on pattern recognition, sequential Markov models or be rule-based. It could also possibly involve machine learning techniques. We should note that there is no absolute distinction between computer-based automatic analysis and computer-based manual analysis. Rather, there are many forms of partially manual and partially automatic analysis. Thus, we may speak of Computer Aided Manual Analysis (CAMA) and Manually Aided Computer Analysis (MACA).

A third relevant issue concerns whether our analysis should aim for representativity or not. Obviously a multimodal corpus can be used as a basis for one or more case studies and since analysis of multimodal data is often very time-consuming and therefore costly, in the past recorded multimodal material have often primarily resulted in case studies. This is completely acceptable and often what is needed in the initial stages of development of a field of inquiry. However, it might also be said that it does not fully exploit the potential which exists in a corpus. This potential is mostly made more full use of by a series of case studies or by a more frequency and statistics-based analysis of patterns in the data than by a single case study.

### *3.7 What can we analyze?*

Finally, we come to the question of what we can analyze in a multimodal corpus. Obviously, this question has no definite answer, since the limits are set by the creativity and insight of the researchers who are doing the analysis. All we can do is point to some examples where multimodal corpora have been used or could be used. The examples will be grouped in three areas, (i) human-human face-to-face communication, (ii) media of communication, (iii) applications.

1. Human-human face-to-face communication.

In many ways this area is the basic area of investigation for multimodal communication. The following are some of the topics that can be investigated in the area. The list is in no way exhaustive.

(i) The nature of communicative gestures. Work has been done on basic taxonomies of both the content/function of gestures and the particular types of expressive behaviour which is employed cf. Ekman and Friesen (1969) Hjortsjö (1969) or Peirce (1955) for a more general semiotic classification of gestures.

(ii) The nature of multimodal communication with a particular communicative function like feedback (cf. Allwood/Cerrato, 2003), own communication management (Allwood/Nivre/Ahlsén, 1990) or symbolic gestures (Poggi, 2002) hesitation: word finding.

(iii) The relations between gestures of different types, resulting for example in questions like: to what extent are feedback gestures symbolic and to what extent are they iconic or indexical? Symbolic feedback gestures here usually involve head movements for *yes* and *no* that can vary from one cultural and linguistic area to another. There are two main variants for *yes* (i) nodding (most European influenced cultures) and (ii) sideways movement of the head back and forth (India). Similarly, there are two main variants for *no* (i) shaking (most European influenced cultures) and (ii) backwards head jerk (the Balkan area, Turkey and the Middle East). There is also iconic feedback, as when one communicator repeats or imitates the movements of another in order to indicate (or display) consensual coordination or agreement. Finally, there is indexical feedback which can be given, for example, by facial gestures, indicating or displaying emotions and attitudes.
A more specific inquiry might here concern when and where in relation to the interlocutor's communication, the different kinds of feedback are used, e.g. when do communicators indicate friendliness by a smile and when do they signal it symbolically by a phrase *like I am so happy to see you*? When are both means used to support each other and when is only one means sufficient?

(iv) The issues raised in (iii) fairly directly lead to the more general question of what kind of relations hold between speech and gestures. What information is usually spoken and what information is usually gestured? How do the two types of information influence each other? How is information distributed on these two modes of production and how is it integrated by us in interpreting other people's contributions to communication. As an extension of this, we can also study the relationship between text and picture in illustrated written text. What information is contributed by text and what information by pictures? How do the modes of representation influence each other (see below)?

Another issue here concerns the temporal relation between spoken utterances and gestures with related content. Is there a fixed temporal order, so that the gestures always come before, after or simultaneously with the related words, or is it possibly the case that the order depends on the circumstances of communication in such a way that there is no fixed order? Psychologists like Goldman-Eisler (1968) and Beattie (2003) have claimed that gestures precede speech, while other psychologists, like McNeill (1979), have claimed that they are simultaneous with speech. The issue is not settled and it is likely that more studies of naturalized multimodal corpora will be helpful in deciding.

(v) Multimodal communication in different social activities. How are speech and gestures used for political rhetoric? Compare, for example, a speech in a TV studio versus in front of a large crowd. How are speech and gestures used in relaxed talk or, for example, in conducting auctions?

(vi) Multimodal communication in different national/ethnic cultures. What multimodal differences exist, for example, between two Swedes, two Chinese and two Italians quarrelling or flirting publicly (cf. Allwood, 1985)?

(vii) Communication and consciousness/awareness. To what extent do different components of multimodal communication reveal differences in degrees of awareness/consciousness regarding what we are communicating about (cf. Allwood, 2002)?

(viii) Communication and emotion. One of the most important functions of gestures in communication is to express emotion. How is this done by different types of people in different activities, in different national/ethnic cultures?

(ix) Communication and power. How is power expressed multimodally? Is it true that more powerful people (cf. Mehrabian 1972) have larger and more powerful gestures and that less powerful people have smaller and more constrained gestures?

(x) The relation between primarily communicative and primarily non-communicative action. Mostly when communication occurs at the service of a practical activity, there is a mixture of action which is not primarily communicative with action that is primarily communicative, e.g. a shop assistant silently hands a customer a product or some change (money) and the customer bows and says thank you, etc.

(xi) Differences between persons belonging to different genders, age groups, social classes, regional groups are not only constructed and/or expressed through spoken language but are constructed and/or expressed just as much through gestures and clothes.

(xii) How is multimodal integration in general (often also called “fusion”) achieved in perception and understanding of communication? How do we integrate visual and auditive information in order to arrive at an integrated audio/visual interpretation of what is being communicated?

(xiii) How is multimodal distribution in general (often also called “fission”) achieved in communicative production? What information is expressed through words, through prosody or through gestures?

1. Multimodality in relation to various media of communication

There are also a large number of issues pertaining to the use of more than one modality in relation to media of communication other than those used in face-to-face communication.

(i) Multimodality in writing (books, magazines, newspapers, advertising). A lot more work can be done here, but see some interesting studies by Kress and Leeuwen (2001) and Halliday (1978).

(ii) Multimodality in films: in principle, many of the topics suggested above for face-to-face communication can also be studied in films with attention to the extra (aesthetic) dimensions added to capture an audience.

(iii) Multimodality in songs and music. Performance and experience of music are multimodal. What we see and what we hear influence each other. This can clearly be seen in opera and rock videos but to some extent in all music.

(iv) Multimodality in visual art and sculpture. Disciplines like art history are already developing corpora and data bases of paintings. Probably some of the analytical tools developed here could also be used in studies of face-to-face communication.

## 4. Applications of multimodal communication

Multimodal corpora can provide useful resources in the development of many different computer-based applications, supporting or extending our possibilities to communicate.

(i) Better modes of multimodal human-computer communication (or more generally human machine communication). See the discussion of embodied conversational agents and avatars above.

(ii) Better computer support for multimodal human-human communication. See the discussion above and the contributions discussed in the AMI project (http://amiproject.org).

(iii) Better modes of multimodal communication for persons who are physically challenged (handicapped).

(iv) Better modes of multimodal presentation of information from databases, for example for information extraction or for summarization.

(v) Better multimodal modes of translation and interpretation. For example, when will we get a system that takes as input a person speaking Italian with Italian gestures and gives as output the same person speaking Japanese with Japanese gestures?

(vi) Better modes of multimodal distance language teaching (including gestures).

(vii) Better modes of multimodal distance teaching (and instruction) in general.

(viii) Better multimodal modes of buying and selling (over the internet, object presentation in shops, etc.).

(ix) Computerized multimodal corpora can, of course, also be useful outside of the areas of computer-based applications. In general, they can provide a basis for the study of any kind of communicative behavior in order to fine-tune and improve this behavior. This could, for example, apply to areas like public oratory or presentation techniques, but also to any kind of service or teaching related communication, like doctor-patient communication, lawyer-client communication, teacher-student student etc.

The list can be made much longer and the preceding topics are mostly intended as a pointer to some of the many possibilities.

## 5. Concluding Remarks

This paper has discussed what might be meant by a digitized multimodal corpus and presented some of the factors that might be relevant in establishing multimodal corpora. I have also presented some of the possible research objectives where multimodal corpora could play an instrumental role.

In doing this, I hope to have provided support for the growing realization that if human (linguistic) communication is basically multimodal (which it seems to be from a phylogenetic, ontogenetic and interactive dynamic perspective, then it also requires us to study language and communication multimodally. This, in turn, means that the creation, maintenance and use of multimodal corpora will remain a very important part of the research agenda for studies of language and communication in the future.

## References

All URLs were accessed in January, 2007

Allwood, J. (1985). Intercultural Communication. *Tvärkulturell Kommunikation*. J. Allwood (ed.). Göteborg, Department of Linguistics, Göteborg University. **12**.

Allwood, J. (2001a). Cooperation and Flexibility in Multimodal Communication. *Lecture Notes in Computer Science*. H. Bunt and R.-J. Beun. Berlin/Heidelberg, Springer Verlag. 2155.

Allwood, Jens (2001b). Dialog Coding – Function and Grammar: Göteborg Coding Schemas. *Gothenburg Papers in Theoretical Linguistics*, GPTL 85. Dept. of Linguistics, University of Göteborg, pp.1-67.

Allwood, J. (2002). Bodily Communication - Dimensions of Expression and Content. *Multimodality in Language and Speech Systems*. D. H. I. K. B. Granström. Dordrecht, Kluwer Academic Publishers.

Allwood, J., Björnberg, M., Grönqvist, L., Ahlsén, E., & Ottesjö, C.(2000). The Spoken Language Corpus at the Dept of Linguistics, Göteborg University. *FQS - Forum Qualitative Social Research*, Vol. 1, No. 3. - Dec. 2000, pp 22.

Allwood, Nivre and Ahlsén (1990). Speech Management: On the Non-Written Life of

Speech. *Gothenburg Papers in Theoretical Linguistics*, 58. University of Göteborg, Dept of Linguistics. .Also in *Nordic Journal of Linguistics*, 13, pp. 3-48.

Allwood, J. and L. Cerrato (2003). *A Study of Gestural Feedback Expressions*. First Nordic Symposium on Multimodal Communication, Copenhagen.

Allwood, J., Cerrato, L., Dybkjaer, Jokinen, K., Navaretta, C., and Paggio, P. 2005. The MUMIN Multimodal Coding Scheme. In *NorFa Yearbook 2005*.

Argyle, M. (1988). *Bodily communication*. London: Methuen.

Beattie, G. (2003). *Visible Thought: The New Psychology Of Body Language*. Routledge: London.

Bellugi, U. (1972). Studies in Sign Language. *Psycholinguistics and Total Communication: The State of the Art.* T. J. O'Rourke, Silver Spring**:** 68-84.

Birdwhistell, R. (1952). *Introduction to Kinesics*, University of Louisville Press.

Brazil, D. 81985). The Communicative Value of Intonation in English. E. L. R. University of Birmingham.

Brunswick, E. (1969). *The Conceptual Framework of Psychology*. Chicago, University of Chicago Press.

Cassell, J. (ed.) 2000- Embodied Conversational Agents. Cambridge, Mass.: The MIT Press.

Ekman, P. and W. Friesen (1969). "The Reportoire of Nonverbal Behavior: Categories, Origins, Usage and Coding." *Semiotica* (1): 49-98.

Goldman–Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press.

Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers*. New York, Academic P.

Gratch, J., Mao, W. & Marcella, S. 2006. Modeling social emotions and social attributions. In Run Sun (ed.) *Cognitive Modeling and Multi-Agent Interaction*. Cambridge: Cambridge University Press. pp- 219-251.

Halliday, M. (1978). *Language as Social Semiotic*. London, Edward Arnold.

Hjortsjö, C. H. (1969). *Människans ansikte och mimiska språket*. Malmö, Studentlitteratur.

Jefferson, G. (1984). Transcript Notation. *Structures of Social Action: Studies in Conversation Analysis*. J. M. Atkinson and J. Heritage. Cambridge, Cambridge University Press.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kita, S, Gijn, I. van, and Hulst, H. van der. 1997. Movement Phases in Signs and Co Speech Gestures, and Their Transcription by Human Codes. Paper presented at ,*Gesture and Sign Language in Human-Computer Interaction*, Bielefeld, Germany.

Kress, G. and T. v. Leeuwen (2001). "Multimodal Discourse - The Modes and Media of Contemporary Communication."

Laban, R. (1974). *The Mastery of Movement*. London, Macdonald & Evans.

Lewis, C. T. S., C., Ed. (1966). *A Latin Dictionary*. Oxford, Clarendon Press.

Martin, J.-C., Pelachaud, C., Abrilian, S., Devillers, L., Lamolle & M., Mancini, M. (2005). Levels of representation in the annotation of emotion for the specification of emotion in ecas. *Proceedings of Intelligent Virtual Agents 2005*.

Mc Enery, T. and A. Wilson (2001). *Corpus Linguistics: An Introduction*. Edinburgh, Edinburgh University Press.

McNeill, David (1979). *The Conceptual Basis of Language*. Lawrence Erlbaum, Hillsdale.

Mehrabian, A. (1972). "Nonverbal Communication."

Månsson, A.-C. (2003). *The Relation between Gestures and Semantic Processes*. Göteborg, Department of Linguistics, Göteborg University, Sweden.

Nelfelt, K. (1998). *Simultaneous Sign and Speech: A Multimodal Perspective on the Communication of Hearing-Impaired Children*. Göteborg, Department of Linguistics, Göteborg University.

Nijholt, A., op den Akker, H.J.A. & Heylen, D.K.J. (2006a). *Meetings and Meeting Modeling in Smart Environments, AI and Society, The Journal of Human-Centred Systems*, 20(2):202-220.

Niljholt, A, Rienks, R.J., Zwiers, J, & Reidsma, D. (2006b).Online and Off-line Visualization of Meeting Information and Meeting Support. *The Visual Computer*, Springer, Berlin, Heidelberg, 22(12):965-976.

Nivre, J., Allwood, J., Grönqvist, L., Gunnarsson, M., Ahlsén, E., Vappula, H., Hagman, J., Larsson, S., Sofkova, S., and Ottesjö, C. 2004. *Göteborg Transcription Standard v6.4.*: Department of Linguistics, Göteborg University.

Peirce, C. S. (1955). *Philosophical Writings of Pierce*. J. Buchler. New York, Cover.

Poggi, I. (2002). "Symbolic Gestures: The Case of the Italian Gestionary." *Gesture* **1**: 71-98.

Prillwitz, S. et al. (1989). *HamNoSys. Version 2.0; Hamburger Notationssystem für Gebärdensprache. Eine Einführung*. Internationale Arbeiten zur Gebärdensprache und Kommunikation Gehörloser; 6. Hamburg : Signum .

Stokoe, W. C. (1978). "Sign Language versus Spoken Language." *Sign Language Studies*(18): 69-90.

Svartvik, J. & Quirk, R. (1980). A Corpus of English Conversation. Lund: Gleerups

Zhang, Z., Potamianos, D., Liu, M. & Huang, T.S. (2006). "Robust multi-view multi-camera face detection inside smart rooms using spatio-temporal dynamic programming", *In: Proceedings of Int. Conf. Face Gesture Recog. (FG)*, Southampton, United Kingdom,

## Transcription systems

IPA – International Phonetic Alphabet http://www2.arts.gla.ac.uk/IPA/ipa.html
SAMPA http:// www.phon.ucl.ac.uk/home/**sampa**/home.htm
CA – transcription.Cf. e g. Jefferson, 1984 above
GTS Göteborg Transcriptions Standard
http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3 (See Nivre et al. above)

## Tools of Multimodal Analysis

MULTITOOL www.ling.gu.se/projekt/tal/multitool
ANVIL http://www.dfki.de/~kipp/anvil/
WAVESURFER http://www.speech.kth.se/wavesurfer
NITE/MATE
http://nite.nis.sdu.dk/
http://mate.nis.sdu.dk/

## Network and project home sites

http://www.amiproject.org (AMI)
http://chil.server.de/servlet/is/101 (CHIL)
http://emotion-research.net (HUMAINE)

Jens Allwood, Göteborg