

# Cross-Linguistic Studies of Multimodal Communication

**P. Wittenburg, S. Kita, H. Brugman**

Max-Planck-Institute for Psycholinguistics  
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands  
peter.wittenburg@mpi.nl

## Abstract

Gestures are culture specific forms of arm movements which are used in communication to transfer information to the listener, to guide the planning of the speech production process and to disambiguate the incoming speech. To understand the underlying mechanisms gestures have to be analyzed in cross-linguistic processes. Large projects are necessary covering speakers from various cultural background and many recordings. Such projects can only be successfully carried out, when suitable gesture encoding schemes, generic annotation schemes, powerful tools supporting the schemes and efficient methods for easy resource discovery and management are available. At the Max-Planck-Institute all aspects were tackled.

## 1. Introduction

The MPI for Psycholinguistics has a long history of research on the synchronization between different modalities in human communication. In the 1980s eyetracking signals and signals about pointing gestures produced important information about the mental processes responsible for speech production [1, 2]. Such signals were typically recorded in relation to spoken utterances. The equipment used was designed to make automatic fine grained temporal analysis possible. For gesture registration IR-light based methods were used. More recently, ultrasonic equipment was used for this purpose identifying the location of maximally 8 sources. This tradition is still continued in the baby labs where eye tracking is recorded to study, for example, the focus of childrens' attention during linguistic tasks. In recent years brain imaging methods (EEG, MEG, PET, MRI) have often been added to get online information about brain activities during speech production and perception task.

In the last few years, research using multimodality shifted towards observational methods in communicative situations of various sorts. Child-caretaker interaction is studied with the help of extensive video recordings to better understand how childrens' language learning is influenced by input and environmental factors. The use of various types of gestures (pointing, iconic and emblematic) is studied in different situations. The following studies should be mentioned in particular: (1) ethnography of pointing gestures; (2) gestural facilitation of speaking or understanding; (3) gestural expression of motion events; (4) speech dysfluencies and gestures; (5) influence of gestures on recipients' gaze movement; (5) hemispheric specialization of types of gestures [3, 4, 5, 6, 7, 8]. In addition, studies about sign language and their comparison to gestural patterns were carried out. The goal of these recordings is fundamental research about the relation between language and thought and the role of gesture in human communication. Since gestures are very much dependent on language and culture, most of the recordings are cross-linguistic, i.e. various countries and cultures are included.

Nowadays the study of multimodal communication based on video recordings is much easier. Information

technology allows science to work with digitized video greatly facilitating the analysis work. For the last two years, all recordings at the MPI have been digitized, yielding an online multimedia corpus consisting of more than 7000 sessions (units of linguistic analysis). Gesture studies form a substantial part of these recordings. Powerful corpus management with the help of metadata descriptions and multimodal annotation tools were developed at the institute to enable the type of research explained. Annotations are stored in well-documented formats well adapted to capturing the complexity of the annotation which are typical of multimodal studies.

## 2. Multimodality Research

Multi-modal records allow us not only to approach old research problems in new ways, but also open up entirely new avenues of research. An old issue, for example, is just how 'modular' language processing is, that is to what extent non-linguistic processes can intervene in the course of linguistic processing. This can be studied by looking at the interaction between two entirely different behaviour streams, gesture and speech. A large multi-media corpus of natural dialogue shows, for example, that when speakers self-edit speech, gesture inhibition actually occurs earlier, suggesting interaction between the speech and gesture execution systems. Similarly, in the comprehension process it can be shown that gesture content is incorporated into the immediate 'message'. Eye-tracking shows that speakers can manipulate the likelihood of this by looking at their own gestures, which are then more often fixated by listeners. More fundamentally, we can look at the role of the two cerebral hemispheres in the production of the two behaviour streams, speech and gesture. Careful studies of the gestures of split-brain patients show that gesture production is largely driven from the right hemisphere, while language of course is normally processed in the left.

In addition to contributing to such long-standing theoretical issues, annotated multimedia records also make possible entirely new lines of research. For example, we have been interested in whether the semantic character of a specific language leads to a special construal of a scene to be described. The study of gesture during online production shows that the way a language 'packages' information has a demonstrable effect on the depiction of

a scene in gestures. Turkish for example packages movement with direction in a single clause but puts manner of motion into a separate adverbial clause ('The ball descended, rolling') – while English allows manner and direction to occur in the same simple clause ('The ball rolled down'). Turkish speakers tend to produce separate gestures for direction and manner, while English speakers tend to fuse them. In a similar way, we have been able to study spatial thinking as it occurs in non-spatial domains, by examining the gestures of speakers talking about e.g. kinship relations.

Sign languages are another domain which has been opened up by multi-media technology. Sign languages are fully-expressive languages which utilize not only the hands, but also the face, gaze and even body-posture to construct complex utterances with phonology, morphology, syntax and 'prosody'. These different 'articulators' express different distinctions in overlapping time windows, where the offset can indicate e.g. the scope of a question. Even the simplest description of a signed utterance therefore requires a multi-tiered annotation of a video-record, and the development of such annotation tools make possible systematic databases for sign language research for the first time. Fascinating questions can now be pursued about effects of modality on language – for example does the spatial nature of the visual-gestural channel have profound effects on the nature of sign languages, and give sign languages an underlying commonality? Most deaf signers are exposed to the gestural systems of the surrounding spoken language, and we can also ask to what extent these gestural systems are recruited into the sign language. Preliminary results from the study of a sign language in the process of standardization (Nicaraguan sign language) suggests that there is such an interaction.

These examples should serve to indicate just what a revolution in our understanding of language and its relation to other aspects of cognition is being made possible by the new technologies. There are also fundamental advantages to archiving multi-media records for all branches of the language sciences. For example, studies of the acquisition of language are hugely enriched by having available the very scene available to the infant language user – we now know for example that unexpressed arguments (e.g. subjects and objects) in Inuit care-takers' speech are often recoverable by the child just because they are most likely in the child's field of view at the moment of utterance. Similarly, records of dying or endangered languages are greatly enhanced by having visual information correlate with the language use. In all these cases, richly annotated multi-media records make possible the extraction of systematic information about the correlation of linguistic and non-linguistic events.

### 3. Gesture Encoding Schemes

#### General

This variety of studies all based on observational methods (i.e. audio and video, sometimes also gaze) required many different gesture encoding schemes on the different

linguistic levels, efficient procedures and powerful tools. Since our researchers are involved in international projects broad agreements on the methods for encoding multimodal behavior are very important. Yet for international standards it seems to be too early, the discipline is too young, although it would facilitate integrating and comparing the data of all the scholarly work.

Most of the studies require careful encoding of the articulator movements<sup>1</sup> and their global timing pattern. Naturally, we are faced with similar problems to those for identifying the articulator movements in the case of speech production. The articulator movements form a continuum, are overlapping and have tolerances dependent on the situation. Therefore, it is not only difficult to make proper time segmentation, but also to classify them.

For gestures which are movements of the arms and its parts accompanying verbal communication acts, it is sufficient to annotate their type and meaning in addition to the articulators. The type of a gesture is a taxonomic classification of its principle purpose and role in communication. It is widely accepted to separate between pointing, iconic and emblematic gestures. Pointing gestures refer to a spatial point or a movement. They appear either as isolated gestures where the meaning is obvious to the listener or mostly in overlap with verbal utterances where the gestures are much more simple to generate and interpret than verbal descriptions. Their meaning is easy to describe by the object they refer to and their intrinsic purpose. Also iconic gestures appear spontaneously as co-speech activities while emblematic gestures stand alone. Iconic gestures have a culturally bound meaning since they are widely accepted within an area.

Gestures often correlate with emotional state, are used to facilitate the planning of speech production and to facilitate speech perception due to their disambiguation capability. Emotional state can be described, although there are no clear conventions yet.

#### Articulators in Gestures

The basis of all scientific work when studying gestures is an encoding scheme for the articulator movements. It was soon perceived that an exhaustive gesture encoding including all relevant characteristics would be ideal but impossible (except for small segments). On the other hand the recordings were perceived as so valuable that re-usage for various research questions was anticipated. To cope with this contradiction it was realised that only an iterative encoding approach would suffice where the needs of primary research projects do not hinder the addition of gesture encodings dedicated to completely different research interests. To support research, the underlying

---

<sup>1</sup> For gestures we have as articulators the arms and its parts up to the fingers. Characteristic movements of the head and the eyes in communicative situations are not treated as part of the gesture although they have similar purposes.

scheme should be exhaustive to define a grid allowing easy computational comparison. Therefore, for a number of recordings focused on in the Institute's gesture project, a thorough study was carried out to attain a general gesture encoding scheme that would allow comparative analysis to be made easily.

Based on Kendon's work a more accurate scheme was developed by v. Gijn, vd Hulst and Kita [9] to separate various phases in a gesture. A *MovementUnit* therefore can exist of several *MovementPhrases*. Basically, each of these can be seen as a sequence of a *Preparation* phase, an *ExpressivePhase* and a *Retraction* phase. An *ExpressivePhase* which covers the meaningful nucleus of a gesture is either an *IndependentHold* or a sequence of a *DependentHold*, a *Stroke*, and another *DependentHold*.

*MovementUnit* = *MovementPhrase*\*  
*MovementPhrase* = (*Preparation*) => *ExpressivePhase* => (*Retraction*)  
*ExpressivePhase* = *IndependentHold*  
*ExpressivePhase* = (*DependentHold*) => *Stroke* => (*DependentHold*)  
*Preparation* = (*LiberatingMovement*) => *LocationPreparation* >>  
*HandInternalPreparation*  
*Retraction* (if subsequent movement) = *PartialRetraction*  
= consists of, \* one or several, => discrete transition, () optional,  
>> normally blended out, occasionally discrete transition

The authors developed a set of descriptive criteria to identify the phases and their usefulness was shown in several studies which were successfully annotated by student assistants.

v. Gijn, vd Hulst and Kita also developed an encoding scheme to describe mainly the articulator movements in the *ExpressivePhase* [10]. It is this phase where annotators are confronted with all the about 60 degrees of freedom and where not only the location and shape has to be described but also for example changes in motion and direction. The following aspects are described: *PathMovementShape* (*straight, circle, round, iconic, 7-form, ?-form, x-form, +-form, z-form*), *PathMovement Direction* (*[up/down], [front/back], [ipsilateral/contralateral]*), *HandOrientationChange* (*[supination/pronation], rotation, [flexion | extension], nodding, [ulnar flexion/radial flexion], lateral flexion*), *HandShape Change* (*[opening | closing], [abduction | adduction], [hinging | dehinging], [clawing | declawing], wiggling, opening wave, closing wave, rubbing, cutting*), *HandOrientation* (*[up/down], [front | back], [ipsilateral/contralateral]*), and *HandShape*. For the latter basically the HamNoSys scheme was re-used.

To support the various gesture related research activities simple encoding schemes are most often derived from this exhaustive scheme. The reference back to the unified exhaustive scheme together with the online availability of the annotated multimedia document allows easy re-usage and an enhancement of the annotations. This can either be corrections of the existing or the addition of new tiers.

When encoding gestures it is of great importance to understand the exact time relationships with the verbal utterances. This is not part of the gesture annotation scheme, but the annotation structure scheme has to provide adequate mechanisms.

#### 4. Annotation Structures

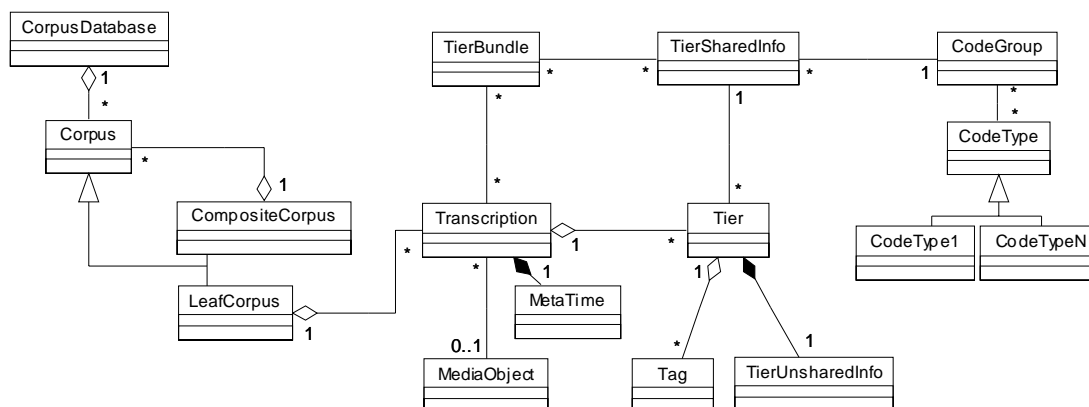
While the encoding scheme describes how to encode the linguistic phenomena (a close handshake in gestures is encoded as "close"), the annotation structure scheme describes the expressive power in structural respects. It has to provide mechanisms for all possible structural phenomena. From our long experience with gesture and sign language studies we know that the annotations can become very complex. There are projects which try to solve this complexity by merging the annotations associated with different linguistic levels into one tier. This method, which is known especially from traditional annotation schemes such as CHAT [11], is also used in new projects. The resulting annotation includes many relations implicitly, i.e. it is the tool which has to include all the knowledge. At the MPI this method was not seen as useful for the future. Different linguistic levels should be separated and all relations such as interruptions, parallelism, semantic correlation should be made explicit.

This is the only way to easily modify the coding later.

In many cases different linguistic interpretations of a gesture are possible. The annotation scheme has to take this into account. Essentially, we follow the indicated way: add another tier which can be used by a new annotator. If only adaptations of the existing annotations are intended, a copy action may be useful for bootstrapping the tier.

The structural phenomena which can occur in annotations are described in detail in [12]. We can summarize the main points:

- The number of tiers can become comparatively high and cannot be seen in advance. It will increase due to various annotators and due to new research goals which require additional information.
- There are all kinds of temporal relations between gesture components and especially between annotations associated with different streams like gestures, speech, facial expression, gaze and others. The complexity makes it necessary to link annotations to periods of time and not to encode overlap and other phenomena in the annotations as older formats require.
- In some occasions spatial relations have to be encoded. They can be encoded as other annotations, i.e. individual or group of coordinate pairs can be linked to time periods.
- In many types of annotations hierarchical relationships have to be included to express linguistic phenomena. These can be token or type oriented. Type specific dependencies are defined at the level of



tier type definitions. Token specific dependencies occur randomly and are defined per linguistic unit.

- Cross-references are very relevant in many cases of linguistic annotation. They describe certain relations which the user wants to draw between two different linguistic units which can be on the same tier or on a completely different one. Comments on some annotation can be interpreted as such cross-reference.

## 5. Abstract Corpus Model

To design the Abstract Corpus Model informal use-case driven method was chosen. In addition a number of existing and well-known annotation formats were analyzed and discussions with linguists about their requirements were carried out. The resulting model defined in UML is more of an operational model than a mere data model.

ACM is realized in first instance as a set of abstract classes that implement common behavior. These abstract classes each have concrete subclasses, one for each of the annotation file formats that ACM currently supports (CHAT, Shoebox [13], relational database [14], Tipster [15], several varieties of XML).

The method calls from ACM's interfaces can be used by a range of annotation related tools. The interfaces are uniform to the tools although the actual objects that implement those interfaces may be instantiated from differently formatted files or even from a relational database. For example, the tools are not aware whether they work on a CHAT file or on a set of database records. Most ACM objects are implemented as remote objects using Java's RMI facilities (Remote Method Invocation). This means that these objects can exist on a central annotation server while the annotation related tools that use their services run on local clients on the network. Method calls to a set of remote interfaces, with arguments and return value, offer a natural way to organize protocols for an annotation server. This type of support for remote objects is efficient since only data that is asked for is sent over the network, i.e. a tier name instead of a complete tier or annotation document. It also forms the basis for a collaborative annotation environment since remote objects can be simultaneously accessed by multiple users. For a class diagram of the first generation of the ACM see figure 1.

It is not the intention of this paper to discuss the part of the class diagram depicted in figure 1 in detail. For this we refer to [16]. But an example can demonstrate how to read it. In this version of ACM, *Tags* have begin and end times that can be specified or unspecified. To make this possible the order of all unaligned *Tags* (i.e. tags which have no specified time marks yet) in a *Transcription* has to be stored explicitly. The object responsible for this is called *MetaTime* and is associated with *Transcription*.

### ACM Revision

Recently, the ACM was revised considerably to include new features. Merging the more elaborated BC (BrowsableCorpus) [17] and EUDICO models of corpora required the introduction of a Session class in ACM. The direct association between Transcriptions and MediaObjects is now administered by a Session object. The composite Corpus structure in ACM is maintained, but as an alternative to BC Corpus hierarchies. There was also a need to introduce Metadata, MetadataContainer and LanguageResource interfaces into ACM as a way to merge in behavior that is needed for BC.

In the first version of ACM, new objects were usually instantiated by their direct ancestors in the corpus tree e.g. Transcription objects were instantiated from LeafCorpus objects. The exact type of the LeafCorpus determined the exact type of the Transcription to be instantiated. In the case of instantiation of a Transcription from a browser over generic corpus trees (like the BC browser) we needed another way to specify the exact type of the Transcription object, and a separate mechanism for creation of this object has to be available.

We were also confronted with a number of related cases where the issue of specifying type and location, and subsequent instantiation of the proper object played a role. For example, in the case of the Spoken Dutch Corpus, currently all digital audio data is delivered on a number of CDROMs. Pointing at and accessing this data, including prompting for the proper CDROM, can be solved by a similar mechanism. For the same corpus, a variation of stand-off annotation is used for annotation documents, where separate annotation tiers are kept in separate XML files in separate directories. Instantiation of an annotation document requires pointing at and combining of these separate files.

To solve this range of problems a design was finished that makes use of the standard mechanisms that Java offers to deal with URLs. Based on a generalization of URL syntax and content type the required access mechanisms (like login prompt, prompt for media carrier) are triggered automatically and the proper type of object is instantiated. In case of ordinary URLs and content types everything automatically falls back on Java's built-in URL handling.

As said, new projects required more complex relations between annotations than the ACM could deal with in its original form. For example, for the Spoken Dutch Corpus both utterances and individual words can be (but don't have to be) time aligned, and each word can have a number of associated codes on different tiers. The Spoken Dutch Corpus also required support for syntactic trees.

For the DoBeS project a wide range of legacy material has to be incorporated in the archive and the EUDICO based archive software has to be able to cope with that. Much of this data is Shoebox or Shoebox-style MS Word data. Therefore interlinear glossing formats have to be supported at the level of ACM. Within the DoBeS community, the maximal format requirements are well described by Lieb and Drude in their Advanced Glossing paper [18].

To support all of these structures two basic types of Annotations were added: `AlignableAnnotations` and `ReferenceAnnotations`. While `AlignableAnnotations` has the necessary characteristics to link annotations to time periods, `ReferenceAnnotations` provide the necessary mechanisms to draw relations between annotations independent of their tier.

In almost every annotation system or format the concept of a tier exists as a kind of natural extension of the concept of a database field applied to time-based data. It is an old idea to "put different things in different places". A tier is the place to put similar things. A tier is a group of annotations that all describe the same type of phenomenon, that all share the same metadata attribute values and that are all subject to the same constraints on annotation structures, on annotation content and on time alignment characteristics.

Metadata attributes for example can be a participant, coder, coding quality, or reference to a parent tier. Constraints on annotation structures can be aspects such as that annotations on the tier refer to exactly one associated parent annotation on a parent tier ( $I - n$ ) or that Annotations on the tier must be ordered in time.

Also annotation content can be constrained by for example a specific closed vocabulary and by a range of possible characters such as Unicode IPA. Constraints on time alignment can also be of various sort such as: Annotations on this tier may not overlap in time.

Explicitly including these types of constraints in the ACM makes tool support for a wide range of use cases and for user interface optimizations possible. For example, known begin or end times of annotations can be reused for new annotations or as constraints on the time segment of other annotations. Text entry boxes can be set up automatically

with the proper input method for IPA, annotation values can be specified using popup menus.

Tier metadata, with attribute values specified or not specified, combined with the tier constraints could be reused as a template for the creation and configuration of new tiers, either in the same document or in another. One step further, a set of tier templates could be part of a document template, making it possible to reuse complete configurations of tiers for other documents.

## 6. Interchange Format

A direct consequence of the ACM is the definition of a suitable and powerful enough annotation interchange format. It is seen as a framework allowing to make ACM content persistent. Here our intentions are fairly comparable with what is currently worked out especially at NIST - called the ATLAS Interchange Format (AIF) [19]. Since AIF could not yet handle all necessary requirements (AIF did not yet support a tier concept) a EUDICO Interchange Format was defined (EAF, see Appendix). However, we would like to join the AIF train to achieve a high degree of interoperability world-wide. Its main structural components are: (1) Time slot values referring to as many as needed concrete time values; (2) information about the tier types and (3) as many `AlignableAnnotations` or `ReferenceAnnotations` as necessary. While the first refer to time slots, the latter refers to annotation IDs.

## 7. Tools

To provide researchers with an efficient annotation and analysis environment, the Institute began early on to setup digitization lines and to build true multimedia tools. The first was the MAC-based MediaTagger annotation tool [20] built in 1994. Consequently, the Institute decided to fully rely on all-digital techniques, i.e. all video and audio signals were digitized. For video it was decided to rely on MPEG1 (after an initial phase of using MJPEG and CINEPAK). Due to its limited resolution, for example, to identify facial expressions in field recordings, it was now decided to change to MPEG2 as a basis for the multimedia archive which has a factor of about 3 more data and bandwidth.

The development of the Java-based EUDICO Tool Set for annotating and exploiting multimedia signals was begun in 1998 and has now reached a flexibility and functionality which makes it one of the most advanced tools for multimodal work. Its nucleus is based on ACM, i.e. it has a comprehensive internal representational power. It has a flexible and easy-to-use annotation and time linking component which allows the user to define his tier setup, which can work with audio and/or video signals in the same way and which makes it possible to do the annotation in various writing systems. It has input methods, for example, for IPA, Chinese, Cyrillic, Hebrew and Arabic. Annotations can either be linked to moments in time in the media stream or to other annotations. It is possible to include hierarchical annotations which is necessary, for example, for an interlinearized representation of morphology.

The EUDICO tool set also provides various views on the multimedia data which can be sound, video, or annotation tracks or other types of signals such as eye tracking tracks. There are a number of stereotypic views on the annotations scientists prefer, therefore EUDICO supports different views and more views can be added according to individual scientists' needs. An important feature is that researchers can easily select and arrange the data tracks they want to see. All viewers in EUDICO are synchronized, i.e. whenever the cursor in a viewer is set to a certain time or segment, all other viewers will move to that instance. The tool set also has a flexible search interface which allows the user to define patterns and associate them with annotation tiers (including all supported input methods) making it possible to enter complex patterns covering several tiers and distances between the patterns. The EUDICO tool set can work in a fully distributed environment where annotation and media tracks are at different locations and support media streaming of fragments. An XML-based generic interchange format was defined (EUDICO Annotation Format), but other formats such as rDBMS, CHAT and Shoebox are also supported.

understand the basic mechanisms of the speech production and comprehension processes. Further the usage of gestures in various cultures could help clarifying the relationship between language and thought. Gestures are very much dependent on the culture and the languages spoken in these cultures.

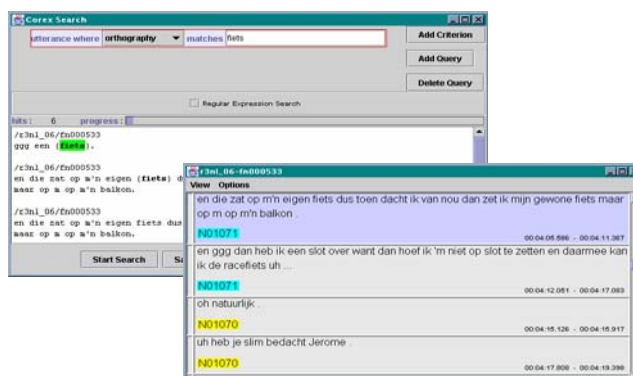


Figure 9 gives an impression of the search feature. It basically allows the user to define search patterns, associate them with tiers and logically combine these patterns to a complete query where also distances can be specified. The result is a list of hits which can be clicked to directly yield the corresponding fragment.

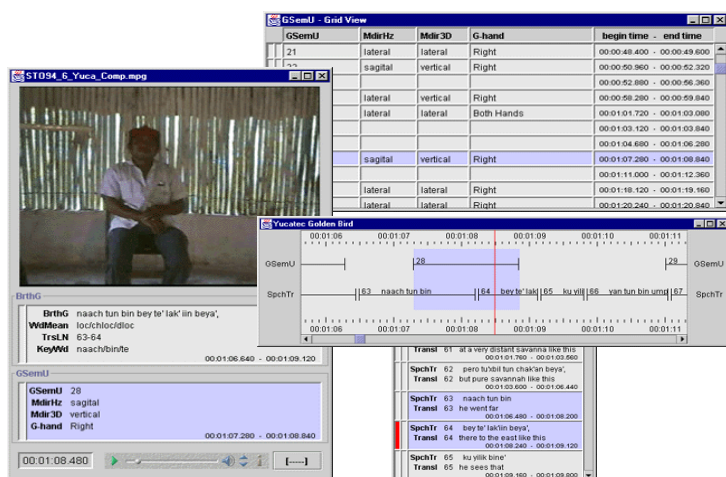


Figure 8 shows the visualization power of EUDICO. Dependent on the project different stereotypic visualizations of the material can be selected. The type of output, the tiers and the order of tiers can be selected by the user. The range of viewers covers dynamic subtitles, a time line view and text viewers with compressed texts.

Tier types can be defined including controlled vocabularies and constraints. Pixel management is very important when dealing with complex tier structures. The user can define the tiers he wants to see and specify the order of presentation. Currently, MPEG1 streaming is supported. MPEG2 is also supported, however downsizing of the video widget is absolutely necessary in order to see the annotations as well.

Further details about the EUDICO Tool Set can be seen on the web-page [21].

## 8. Conclusions

At the MPI for Psycholinguistic the study of gestures has a long tradition. Gesture recordings are used to better

To support this research a large cross-linguistic gesture corpus had to be built including annotations of the speech acts and the gestures. Currently, large international projects have been setup to further investigate the scientific questions raised in this paper.

Such research was only possible by a consequent digitization policy of the institute, by building efficient multimodal annotation and exploitation tools and by powerful mechanisms which help the user to manage large corpora. With the EUDICO and Browsible Corpus technology which was extended within the ISLE project the researchers can rely on tools which will be supported for many years. Since the file formats of both technologies is XML based it can be expected that they will be widely used.

## 9. References

- [1] W.J.M. Levelt (1980). Online processing constraints on the properties of signed and spoken language. In Biological Constraints on linguistic form. U. Bellugi, M. Studdert-Kennedy (eds.). Vgl. Chemie, Weinheim.
- [2] G. Richardson (1984). Word recognition under spatial transformation in retarded and normal readers. Journal of Experimental Child Psychology 38, 220-240.
- [3] S. Kita, J. Essegbey (to appear). Pointing left in Ghana: How a taboo on the use of the left hand influences gestural practice. Gesture.
- [4] S. Kita (1998). Expressing a turn at an invisible location in route direction. In Ernest Hess-Lüttich, J.E. Müller & A. vanZoest (eds.), Signs & SPace. 159-172. Tübingen: Narr.
- [5] A. Özyürek, S. Kita (1999). Expressing manner and path in English and Turkish: Differences in speech, gestures, and conceptualization. In M. Hahn and C. Stones

- (eds.), Proceedings of the 21 st Annual Meeting of the Cognitive Science Society. 507-512. Amsterdam.
- [6] M. Gullberg, K. Holmqvist (2001). Eye tracking and the perception of gestures in face-to-face interaction vs. on screen. In C. Cave, I. Guaitella, S. Santi (Eds.), *Oralite et gesturalite: Interactions et comportemetns multimodaux dans la communication* (pp. 381-384). Paris: L'Harmattan.
- [7] H. Lausberg, S. Kita (2001). Hemispheric specialization in spontaneous gesticulation investigated in split-brain patients. In C. Cave, I. Guaitella, S. Santi (Eds.), *Oralite et gesturalite: Interactions et comportemetns multimodaux dans la communication* (pp. 431-434). Paris: L'Harmattan.
- [8] M. Seyfeddinipur, S. Kita (2001). Gesture and dysfluency in speech. In C. Cave, I. Guaitella, S. Santi (Eds.), *Oralite et gesturalite: Interactions et comportemetns multimodaux dans la communication* (pp. 266-270). Paris: L'Harmattan.
- [9] S. Kita, I. v. Gijn, H. vd. Hulst (1998). Movement Phases in Signs and Co-speech Gestures, and their Transcription by Human Coders. In I. Wachsmuth and Martin Frühlich (eds.), *Gesture and Sign Language in Human-Computer Interaction*, Vol. 1371: 23-35. Proceedings of the International Gesture Workshop Bielefeld, Lecture Notes in Artificial Intelligence. Berlin: Springer Verlag.
- [10] S. Kita, I. v. Gijn, H. vd. Hulst (2000). *Gesture Encoding*. MPI Internal Report.
- [11] B. MacWhinney (1999). *The CHILDES Project: tools for analyzing Talk*. Second ed. Hillsdale, NJ: Lawrence Erlbaum.
- [12] S. Levinson, S. Kita, P. Wittenburg, H. Brugman (2002). Multimodal Annotations in Gesture and Sign Language Studies. In *Proceedings of the LREC 2002 Conference*, Las Palmas.
- [13] [www.sil.org/computing/catalog/shoebbox.html](http://www.sil.org/computing/catalog/shoebbox.html)
- [14] [www.mpi.nl/world/tg/CAVA/CAVA.html](http://www.mpi.nl/world/tg/CAVA/CAVA.html)
- [15] [www.cs.nyu.edu/cs/faculty/grishman/tipster.html](http://www.cs.nyu.edu/cs/faculty/grishman/tipster.html)
- [16] H. Brugman, P. Wittenburg (2001). The application of annotation models for the construction of databases and tools. In *Proceedings of the Workshop on Linguistic Databases*. Philadelphia.
- [17] [www.mpi.nl/ISLE](http://www.mpi.nl/ISLE)
- [18] H. Lieb, S. Drude (2000). *Advanced Glossing: A language documentation format*. Unpublished working paper.
- [29] [www.nist.gov/speech/atlas](http://www.nist.gov/speech/atlas)
- [20] [www.mpi.nl/world/tg/lapp/mt/mt.html](http://www.mpi.nl/world/tg/lapp/mt/mt.html)
- [21] [www.mpi.nl/world/tg/lapp/eudico/eudico.html](http://www.mpi.nl/world/tg/lapp/eudico/eudico.html)  
[www.mpi.nl/tools](http://www.mpi.nl/tools)

## 10. Appendix

This appendix contains the DTD for the EUDICO Annotation Format (EAF).

<!-- edited with XML Spy v4.1 U (<http://www.xmlspy.com>) by Hennie Brugman (Technical Group) -->

<!--  
Eudico Annotation Format DTD  
version 0.1  
July 5, 2001

```
-->
<!ELEMENT ANNOTATION_DOCUMENT (HEADER,
    TIME_ORDER, TIER*, LINGUISTIC_TYPE*, LOCALE*)>
<!ATTLIST ANNOTATION_DOCUMENT
    DATE CDATA #REQUIRED
    AUTHOR CDATA #REQUIRED
    VERSION CDATA #REQUIRED
    FORMAT CDATA #FIXED "1.0"
>
<!ELEMENT HEADER EMPTY>
<!ATTLIST HEADER
    MEDIA_FILE CDATA #REQUIRED
    TIME_UNITS (NTSC-frames | PAL-frames | milliseconds)
    "milliseconds"
>
<!ELEMENT TIME_ORDER (TIME_SLOT*)>
<!ELEMENT TIME_SLOT EMPTY>
<!ATTLIST TIME_SLOT
    TIME_SLOT_ID ID #REQUIRED
    TIME_VALUE CDATA #IMPLIED
>
<!ELEMENT TIER (ANNOTATION*)>
<!ATTLIST TIER
    TIER_ID ID #REQUIRED
    PARTICIPANT CDATA #IMPLIED
    LINGUISTIC_TYPE_REF IDREF #REQUIRED
    DEFAULT_LOCALE IDREF #IMPLIED
    PARENT_REF IDREF #IMPLIED
>
<!ELEMENT ANNOTATION (ALIGNABLE_ANNOTATION |
    REF_ANNOTATION)>
<!ELEMENT ALIGNABLE_ANNOTATION
    (ANNOTATION_VALUE)>
<!ATTLIST ALIGNABLE_ANNOTATION
    ANNOTATION_ID ID #REQUIRED
    TIME_SLOT_REF1 IDREF #REQUIRED
    TIME_SLOT_REF2 IDREF #REQUIRED
>
<!ELEMENT REF_ANNOTATION (ANNOTATION_VALUE)>
<!ATTLIST REF_ANNOTATION
    ANNOTATION_ID ID #REQUIRED
    ANNOTATION_REF IDREF #REQUIRED
    PREVIOUS_ANNOTATION IDREF #IMPLIED
>
<!ELEMENT ANNOTATION_VALUE (#PCDATA)>
<!ELEMENT LINGUISTIC_TYPE EMPTY>
<!ATTLIST LINGUISTIC_TYPE
    LINGUISTIC_TYPE_ID ID #REQUIRED
>
<!ELEMENT LOCALE EMPTY>
<!ATTLIST LOCALE
    LANGUAGE_CODE ID #REQUIRED
    COUNTRY_CODE CDATA #IMPLIED
    VARIANT CDATA #IMPLIED
>
```