

Language Resource Archiving supporting Multimodality Research

Peter Wittenburg, Daan Broeder, Peter Berck, Han Sloetjes, Alex Klassmann
Max-Planck-Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
peter.wittenburg@mpi.nl

Abstract

At the MPI multimodal research has a long history. An increasing amount of resources is created to test scientific hypothesis. This requires proper methods and technologies to manage these resources. During the last five years mature tools¹ were developed for these purposes that guide the resources during their whole life-cycle; ELAN can be used to create accurate and complex annotations; IMDI helps the user to create useful metadata descriptions, to model the underlying relations between the resources and to search for suitable resources; LAMUS is used to upload and manage large language resource repositories and finally ANNEX and LEXUS can be used to access multimodal resources via the web.

Introduction

Investigating multimodal behavior was and is one of the key pillars in psycholinguistic research to get a deeper understanding of the mental processes underlying speech production and speech comprehension, and to better understand the relation between language and cognition. Therefore, at the Max-Planck-Institute for Psycholinguistics many studies were and are carried out using a number of modalities such as speech, prosody, gestures, signs, eye movements and body movements [1-11]. Different recording techniques such as video, audio, eye trackers, data glove, motion trackers, and ultrasonic and infrared marker devices were and are used to gather multimodal data.

As a result of the various research projects carried out at the MPI for Psycholinguistics its language resource archive now covers about 150.000 objects most of which are sessions that are linguistically meaningful units such as interviews, route descriptions, narratives etc. This is covered in about 15 Terabytes of data, a large amount since much data is digitized video. Only video (including audio) signals, sometimes taken from different perspectives, carry the rich information that is necessary to analyze and annotate human communicative behavior. In general the annotation is a manual process since these signals are often

recorded in natural environments and contain utterances in minority languages or spoken by children, second language learners etc. i.e. there are no proper language models, the corpora are in general too small to estimate parameters for stochastic recognition machines and the signals contain too rich information. Signals such as eye tracking data is normally not annotated, but just used to determine relevant points in time where for example the eyes fixate a certain pattern.

A large part of this archived data is well-organized and described with the help of the IMDI (ISLE Metadata Initiative) metadata infrastructure. Although the IMDI set² contains elements that are typical to describe multimodality it is difficult to guess how much of the data in the archive is actually used for multimodality research. In principle, any video recording can be used to carry out such studies and researchers often forget to mark multimodality in the metadata descriptions. Therefore, we can only refer to institute projects that are started purposefully to include multimodal analysis (see annual reports³).

This paper will focus on the aspects of managing the technical complexity that naturally evolves when doing multimodal research, i.e. during annotation, during resource management and during analysis. It will present a framework that allows researchers to carry out multimodality work with a high accuracy and efficiently. It will not focus on either scientific results, models of multimodal behavior in production and comprehension, and encoding schemes that are used to encode human behavior. For further information about the more scientific aspects we refer to the annual reports of the MPI.

Annotation Schemes

Despite some research projects in the area of iconic gestures, stereotypical tasks such as “route description” where speech and gestures are recorded and annotated according to a more general schema [11,12] we cannot speak about the emergence and broad usage of generic schemas for the encoding of linguistic phenomena. It is understood that a bottom up description of

¹ All tools are available or will soon be available under Open Source license. For details we refer to the following web-site: www.mpi.nl/tools

² www.mpi.nl/IMDI

³ www.mpi.nl/research/publications/AnnualReports

multimodal streams starting with articulator movements is enormously complex and therefore not tractable. Instead of that researchers are looking for encoding schemes at the semantic level that allow them directly to test their scientific hypothesis. Therefore, almost all studies invent new schemes to do the linguistic encoding.

However, it was of great importance to define an annotation scheme at the structural level which is powerful enough to represent the linguistically interesting phenomena with the required flexibility and time granularity. Therefore, a.o. the EAF XML (ELAN Annotation Format) schema was developed and improved over time. It allows the researcher to define (and modify) his/her own tier setup to encode the behavior at any linguistically relevant level, to encode dependency relationships between them and to connect them where necessary to the time axis. Although all meaningful behavior is generated by mental processes we cannot speak about dependent streams, multimodal streams such as for example speech, eye movements and gestures have to be treated as completely independent, i.e., all types of timing relationships can occur [13]. Therefore, EAF has the notion of “time references” so that any single annotation can be associated with a period of time on the axis⁴. On the other hand, there will be hierarchical relations such as between the movement of the whole arm and that one of the hand in gestures or between the spoken words included in a verbal utterance. To cope with these phenomena EAF has the notion of “hierarchical relations” that can evolve to trees of different depth during encoding. Often phenomena are not linked to the time axis, but refer to an element on another tier such as in interlinearized representations. To cope with this EAF introduced the notion of “symbolic references”. In linguistic encoding often phenomena are related that are on the same tier but non-adjacent time periods or that are on different tiers. An annotation format therefore has to support the encoding of such phenomena as well.

To normalize the timing encoding we should add that points in time that are used to anchor annotations are shared and stored as ordered sequence. This is in accordance with the Annotation Graph model from Bird and Liberman [14].

The archive still contains many multimodal annotations that were created with the MediaTagger tool [15] which was one of the first

supporting flexible multimodal annotation. However, the underlying model had limitations [16] and it was not compliant with the Annotation Graph model. Due to changes in the internal Quicktime representation and due to operating system peculiarities⁵ the conversion process to the more generic and XML-based EAF format turns out to be a time consuming process. Some limited multimodal annotations are also done by making use of CHAT [17]. Import modules are available to easily convert them into EAF format. With the exception of the MediaTagger resources we can argue that all multimodal annotations are available in archivable formats.

Management and Access Architecture

Given the large amount of (multimedia/multimodal) resources created and stored at the MPI we had to work out an architecture that supports their whole life cycle from creation to usage and long-term preservation. The following figure gives an overview of the architecture and except the web-based annotation creation and commentary all components have been implemented and are in operation.

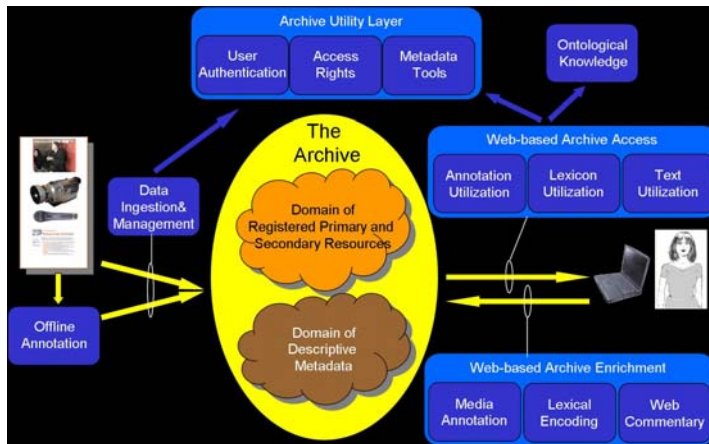
In the following we will briefly describe the architecture and then explain some components in more detail. The user can:

- off-line annotate and analyze recordings by ELAN (not included in the figure)
- describe them with metadata using the IMDI Editor (Metadata Tools)
- upload them into the archive with LAMUS (Data Ingestion and Management)
- define suitable access policies with AMS (User Authentication and Access Rights)
- search and browse for suitable resources with the IMDI tools (local and web-based, Metadata Tools)
- download one or complete sub-archives with the IMDI tools (Metadata Tools)
- carry out content searches on the annotations and visualize the annotated media recordings with ANNEX (Annotation Utilization, Media Annotation)
- manipulate lexica with LEXUS if applicable (Lexical Utilization, Lexical Encoding).

In addition, services take care that several instances of the recordings are stored at different locations in

⁴ It is assumed that preprocessing is used to unify the time axis underlying different recordings. The ELAN tool for example has a few operators to carry out this unification.

⁵ Former MAC-OS version made a difference between data and resource fork information being both crucial for a correct interpretation. However, copying activities were carried out without awareness of this relation.



the Netherlands and Germany to ensure long-term survival.

Creating Multimodal Resources

The latest version of ELAN offers many advanced features that facilitate the time-consuming manual annotation work. It not only supports the flexible annotation model described above, but it also deals with different types of media streams or time series as they occur in multimodal observations. Video signals resulting from several cameras can be displayed and analyzed together with time series data created by the many channels of for example a data glove device. All streams and the created annotations are time synchronized, i.e. selecting a time fragment in one viewer will directly update the position in the other viewers. Different options for visualizing the complex annotations that easily can contain more than 20 layers help users while navigating, comparing instances of similar phenomena etc. In many studies of multimodal interaction precise time accuracy in the order of video-frames is of greatest importance. This is the reason why we asked SPEX, the Dutch center for evaluations, to carry out measurements about the accuracy of ELAN. While earlier versions of ELAN made use of JavaMediaFramework the later versions make use of the native media libraries on the Windows platform. Together with well-chosen MPEG codecs it was shown that this solution offers the required frame accuracy in annotation and in playing. For other research projects where time accuracy is not that important, but where efficiency is the primary criterion ELAN offers a fast tagging mode.

For a detailed description of the features of ELAN we refer to the manual which is available on the web. Now ELAN has reached a level of maturity that it is a tool widely used for multimodal and sign language studies. All annotations are represented in XML which makes it a suitable candidate as well for data that has to be archived.

Managing Resources

Multimodal research is accompanied in general by a large amount of resources that are related in various ways: media recordings from different devices are related since they share the same time axis, annotations are linked with specific recording channels, recording sessions are embedded in experimental setups etc. It is very important to store this relation information. In the MPI setup this is done by using the IMDI framework. IMDI allows users not only to carefully describe the sessions, but also to express the different type of

relations. In doing so metadata descriptions are supporting the user in creating a well-organized browsable archive that can be accessed by searching as well as by browsing. IMDI therefore is the basis for managing a large amount of closely related resources as they are typical for multimodality research. The possibility of fine-grained metadata descriptions can be used to formulate scientifically interesting queries, in particular in conjunction with queries about the content of annotations. Also IMDI files adhere to a publicly available XML schema and therefore are in an archivable format.

The LAMUS system (Language Archive Management and Upload System) is used as a gate keeper for the language resource archive at the MPI. It allows depositors and managers to add new resources to the archive and to update existing ones in a way that the archive remains coherent and consistent. A user can request a workspace for some period of time, define an uplink in the archive as anchor point for new resources, create corpus structure, add metadata descriptions and integrate resources. Once all manipulations in the workspace have been carried out to the satisfaction of the user or after a validation procedure has been taken place, the workspace content can be uploaded to become part of the archive. LAMUS, therefore, controls the consistency of the archive and with the help of configurable resource type files its coherence. The configuration file specifies the accepted formats and when parsers are available checks the formal correctness of the ingested files.

LAMUS has an access management component that allows depositors or managers to define access policies and rights with powerful commands such as "make all audio resources in this sub-corpus available to all". Policies can be defined that determine the kind of declarations (code of conducts etc) users have to accept before getting access to resources that are protected. Another

important aspect is that LAMUS initiates the creation and updating of indexes for fast searching in metadata and annotations whenever a new resource is uploaded.

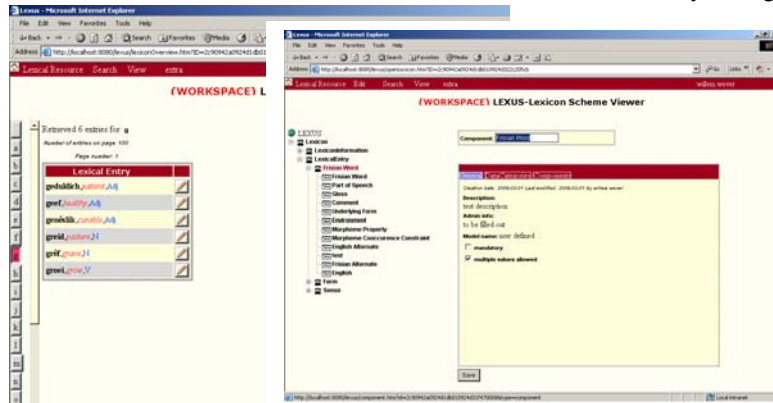
Accessing Resources

The MPI archive provides a number of access methods knowing that researchers have different wishes. The most simple is to browse and search in the metadata domain to find useful resources. Once found they can be downloaded or viewed with a normal plug-in. Many researchers, however, want to carry out analysis on their computer by using own software and therefore want to access and operate on a number of files. They are offered a Tree-Copier option in the metadata browser that allows them to specify a sub-corpus in the archive and download all resources (or only those of specific type). The metadata and corpus structure information is also copied so that the user has a complete local copy of this sub-archive, that can also be browsed and searched using local tools just as the mother archive. Tools can then be used to carry out some manipulations off-line and later, with the help of LAMUS, that part could be uploaded again.

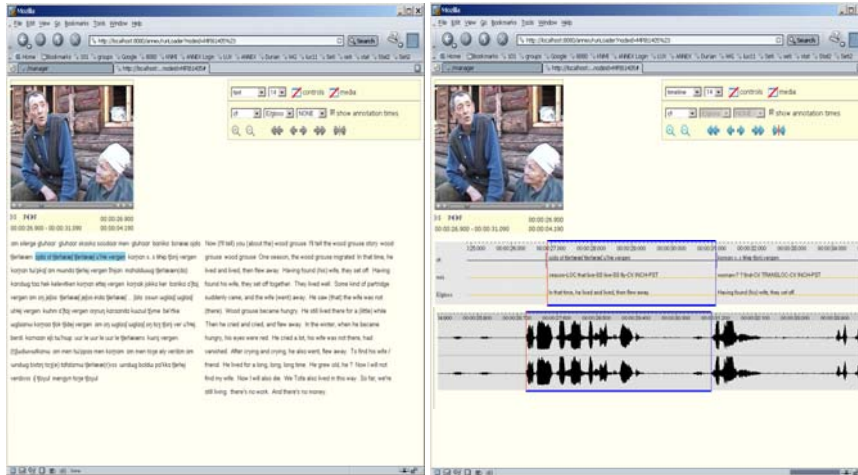
More interesting, however, are web-based applications that allow the user to immediately visualize objects such as multimedia lexical annotations. ANNEX is a flexible tool

media presentations. This is one of the reasons that ANNEX does not yet allow to support annotating. Also ANNEX comes along with synchronized viewers, different viewers for annotations and it offers search capabilities on the annotation content. The following figures give two views of the look and feel of ANNEX. Again the further details can be found in the manual.

Another web-based tool that can be mentioned is the LEXUS lexicon tool which allows the creation and modification of lexical information, i.e. information that is abstracted from the individual occurrence in annotations. Since LEXUS also can incorporate or link to multimedia signals (photos, audios, videos) it can be used for example to store typical signs or gestures. LEXUS is based on LMF (Lexical Markup Framework) which is a generic model currently being



The right part shows a view on the structure and the information about the data categories used in a typical LMF-based lexicon with LEXUS. The left part gives one of the possible views of the content which can be easily browsed for example by selecting the begin character.



Two screen shots indicating the features of ANNEX to visualize and listen to annotated media files.

to work on annotations in a similar way as ELAN does it. Due to the functioning of the web, no guarantees can be made for the smoothness of the

search functionality and many other features and a first interaction with ANNEX has been implemented. The following figures may give an impression of the look and feel of LEXUS. The program offers too many features amongst which is web-based structure and content manipulation and collaboration so that we like to refer to the manual for details.

Conclusions

For the many multimodal research projects carried out at the MPI and in collaboration with other institutions a complete set of technologies was developed that can support the whole life-cycle of multimodal annotations. The lack of agreed linguistic annotation schemes in multimodal research due to the specificity of the research questions increased the necessity to define and apply a flexible annotation format such as EAF and to apply a flexible lexicon format such as LMF (for the rare cases where lexical abstractions are required in multimodality projects). In ever growing corpora with many interrelated resources IMDI is an excellent way to not only create a meaningful organize, but to carry out scientifically relevant searches in combination with searches on the annotation content.

The language resource archive serves as a reliable repository that can be accessed in several ways leaving enough flexibility for the individual researcher. Its dynamic nature and the move towards web-based applications require the introduction of Unique Resource Identifiers and a smart linguistically motivated versioning. Both will be included in the next LAMUS versions. Still many checks have to be added to ensure consistency of the representations at the structural, format and metadata level. But this has to be balanced with the requirement of flexibility. Still many tools generate formats that are not schema-based and therefore difficult to validate.

The MPI will continue to develop its technology and continue to make it available to other interested institutions under Open Source licenses. A first external installation was finished successfully at Lund university, other external setups will follow.

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 comportements 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 comportements 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 comportements multimodaux dans la communication* (pp. 266-270). Paris: L'Harmattan.
- [9] N. Enfield. 2002. Hand pointing in Laos: form and function in a locality description task. *MPI Annual Report 2002*. Nijmegen.
- [10] U. Zeshan. 2004. *Sign Language Typology Project*. *MPI Annual Report 2004*. Nijmegen.
- [11] 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.
- [12] S. Kita, I. v. Gijn, H. vd. Hulst (2000). *Gesture Encoding*. *MPI Internal Report*.
- [13] H. Brugman, P. Wittenburg, St. Levinson, S. Kita (2002), *Multimodal Annotations in Gesture and Sign Language Studies*. *LREC 2002 Conference*. Las Palma, Mai
- [14] S. Bird and M. Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1,2):23-60.
- [15] www.mpi.nl/world/tg/CAVA/mt
- [16] H. Brugman and P. Wittenburg. 2001. The application of annotation models for the construction of databases and tools. *IRCS Workshop on Linguistic Databases*, University of Pennsylvania. Philadelphia.
- [17] <http://childes.pst.cmu.edu>