

Aspects of Modern Multi-modal/Multi-media Corpora Exploitation Environments

Daan Broeder, Hennie Brugman, Peter Wittenburg

Max-Planck Institute for Psycholinguistics
The Netherlands
d.g.broeder@mpi.nl

Abstract

This paper wants to discuss several aspects of multimodal/multimedia language resources such as the use of metadata descriptions for easy location purposes, their collaborative annotation and exploitation via Internet, the generation of synchronized media and text streams in distributed environments, and general annotation formats. These aspects that although they may be discussed independently have to fit together seamlessly to offer users an adequate exploitation environment that is up to the huge amount of data that is available in modern multi-media corpora and is able to exploit fully the current technology advancements.

1. Introduction

In the Max-Planck Institute for Psycholinguistics (MPI) we are faced with new challenges in the area of maintaining multimedia language resources. This is due to the trend in the institute to base linguistic theory on multimedia recordings, to digitize all available material to give direct and immediate access, to get involved in a number of projects such as the establishment of a Gesture database and the Documentation of Endangered Languages (DOBES [3]). In the latter about 20 teams will be carrying out many recordings, annotate and analyse them for 5 years.

We have found it helpful to distinguish metadata from annotation data and treat them as orthogonal dimensions. While metadata in this context is meant to describe the whole language resource, the annotation is a time synchronous description of what is happening and is spoken during a recording. In MPEG7 and TEI [5,6] for example these two dimensions are integrated into one file format. The main reason to distinguish these two dimensions is their treatment. While metadata is mainly used to find and select resources, annotations are used to study the nature of communicative acts of humans.

2. Using metadata descriptions

The use of metadata descriptions for corpus exploitation is not new but it has been used in the form of proprietary formats for a singular corpus only. In our opinion metadata descriptions can be used to easily find and locate suitable language resources in the Internet. In the IMDI (ISLE Metadata Initiative) project we try to reach a consensus within a representative part of the linguistic community about

a standard for such metadata descriptions describing language resources. A machine-readable implementation of this standard will then allow us to build up a searchable and browsable space. Our presentation with respect to metadata is partly based on the work executed within the framework of the international EAGLES/ISLE [1,2] project, on practical work with metadata descriptions at the MPI for Psycholinguistics and on suggestions with respect to metadata within projects such as DOBES [3] and CGN [4].

2.1. A Browsable Corpus Universe

A key issue in the framework we envision as a corpus exploitation environment is that corpora should not stand-alone in the universe. Instead a corpus should be embedded in an infrastructure that enables researchers to locate appropriate resources. We propose to build a universe of metadata descriptions for language resources. The metadata descriptions will incorporate links to one another, thus forming a structure that can be browsed and searched. These metadata descriptions serve the goal of corpus discovery as well as the goal of corpus exploitation.

At the MPI this concept was introduced under the name of Browsable Corpus (BC) to help organize and structure a growing mass of multimedia language resources (LR). The BC scheme depends on the creation of an accompanying metadata description file (MDF) for each individual LR and for every LR bundle, e.g. a transcription, or transcription plus media files for a multimedia corpus. The latter, which are called Session MDFs, contain a variety of information about the LR's and also specify where to find the individual resources. Above this lowest layer of metadata description files that point to the LRs a hierarchy of other meta-descriptions is built up. Each MDF refers to at least one MDF lower in the hierarchy, so that the MDFs form subsets that share certain metadata attributes with the session MDF. Together the MDFs form a pyramid culminating in a top MDF that forms the entry point of the corpus universe (see Figure 1).

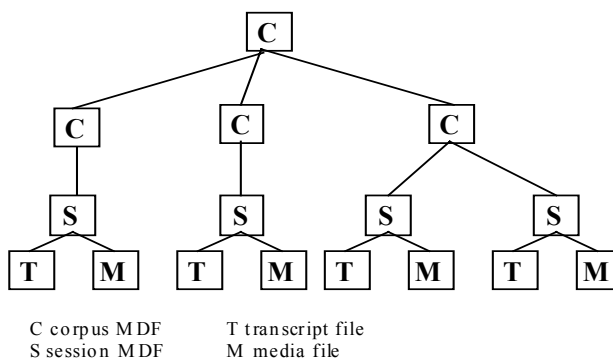


Figure 1 The BC Hierarchy

The structure created by the linked MDFs creates a space that can be navigated by a browser tool specifically designed for this task. At each node within this space the browser will display the metadata for the selected MD node, in a way that is closely analogous to a web browser that shows the content

of the linked HTML pages that form the World Wide Web (see). Since the links between the MDs and LRs are specified as URLs and the browser can access MDFs via the HTTP protocol, the MDFs and LRs may be distributed over different physical sites, reinforcing the WWW analogy.

Browser Action

Exit
About
Options

Status: Ok

History

MPI corpora (UNIX)
MPI Corpora (HTTP)

Meta Descriptions

Xml mpi-unix.xml
 C MPI corpora
 + ESF
 C LAC-corpus Language & Cognition
 + africa
 C australia
 + fareast
 C mesoamerica
 - lac-mesoamerica-yucatec.xml
 C tzeltal
 C Penny Brown
 i Brown-TZELLG.html
 i **Brown-DATATYPE.html**
 i Brown-REFERENC.html
 C Tzeltal Dictionaries
 + Tzeltal Grammar
 + 1971-73 fieldtrip
 + 1980 fieldtrip
 + 1990-2000 fieldtrips
 + zapotec
 - lac-mesoamerica-oltec.xml
 + mideast
 + northamerica
 + oceania

Root URL: http://www-server.mpi.nl/topics/BC/mpix-unix.xml

List Add Remove Remove All

Description

This file gives information about references

Info/Content

```

<HTML>
<HEAD>
<META HTTP-EQUIV="Content-Type" CONTENT="text/html;
charset=windows-1252">
<META NAME="Generator" CONTENT="Microsoft Word 97">
<TITLE>datatype</TITLE>
</HEAD>
<BODY>
<B><FONT SIZE=4><P>Info file describing types of data in the
database:</P>
</B></FONT><FONT SIZE=2><P>&nbsp;</P>
<P>The Tzeltal data in this corpus was collected during three major
periods of fieldwork. The first was P.Brown's PhD dissertation
fieldwork in 1971-73; data consist of audiorecordings, transcripts,
and fieldnotes. The second was a summer of fieldwork in 1980, by
P.Brown and S.C. Levinson, in preparation for (and funded by) a
research project at the Australian National University. The data were
both audiorecorded and (some)8mm film recorded (and later copied
to VHS video). The third period of research is sponsored by the Max
Planck Institute; this started in 1990 and continues through the
present; part of this is in connection with the MPI Space project and
was conducted in collaboration with S.C. Levinson.</P>
<P>&nbsp;</P>
<P>Data collected during the first two periods consists of (i)
naturally-occurring Tzeltal conversation; (ii) Tzeltal speech in public
situations (political speeches, speech at markets, fiestas, church
sermons, court cases), (iii) linguistic elicitation, (iv) elicited songs and
narratives, (v) fieldnotes. The data collected during the decade
1990-2000 includes some of the same kinds of data as above, but
focuses in addition on (vi) spatial language, including
naturally-occurring Tzeltal spatial descriptions in everyday contexts,
in the household, on the trails, in the fields, as well as examples
systematically elicited from both adults and children in response to
  
```

Figure 2 The BC Browser

A nice example of the possibilities if the BC concept was demonstrated during the opening of the “Year of the Language” in Lund earlier this year when examples of (parts of) corpora of six different European institutes were shown linked into a small browsable corpus universe.

2.2. Tools

An initial idea was that the metadata descriptions could also have elements that describe specific tools that can be used to act on the resources themselves. However since resources and tools form orthogonal dimensions it is better to have the metadata description only describe resources and not a set of tools that will change in time anyway. A more elegant solution is to describe the type and structure of the resources in sufficient detail so that “browser” tools used to access the metadata description can decide which ones of the available tools are suitable to handle the data. This is either based on local user configurable information or on a remote tool registry that has knowledge about what tools fit specific resources.

At the moment we are experimenting with a scheme of (semi) mime-types to characterise language resources. We foresee that users will want to customise the mapping of tools to resource types to their own taste just as they are able to do with WWW-browsers.

A special question is the organisation of the infrastructure. Where will we store the metadata descriptions and/or resources? In the IMDI project as preliminary solution we plan to create a registry authority. This registry authority has to build a web-portal, check the quality of the produced meta descriptions, create intuitively understandable browsable hierarchies based on the meta descriptions and link the meta descriptions to other types of information and resources. The registry authority will also provide tools such as a constrained editor that allows the user to create metadata descriptions and a suitable browser which can operate on the metadata description files. The IMDI project will also work on requirements for the registry authority and the metadata tools.

3. Creating and Exploiting Annotations

Modern multimedia resources are characterized by complex annotation structures. This has to do with the facts that multimedia recordings are the basis for annotations of several channels which are treated as being independent from each other and that often cross-references are essential to mark dependencies and relations of various sort. Annotating gesture and speech acts in recordings where several speakers interact can easily lead to more than 20 annotation tiers where the speakers are acting independently and where the speech and gesture annotations have their own partly overlapping relations to time.

Hierarchical relations between the codes within tiers and the wish of the user to mark special relations between several units on different tiers impose new requirements on generic annotation representations. The requirements resulting from various multimedia projects the MPI is taking part in lead to the definition of the Abstract Corpus Model (ACM). This

Object Oriented model was designed such that it can represent operations on all existing annotation structures to be found in current corpora. An implementation of ACM in terms of Java classes is the nucleus of the EUDICO tool set [7], currently being developed at the MPI. With the help of import/export components it is possible to interchange data with relational databases, corpora in the CHAT format [9], corpora in the Shoebox format [8], and data contained in several types of XML-structured files. The ATLAS API and matching ATLAS Interchange Format [10] seem to have similar representational power. We therefore intend to extend EUDICO with a component for data interchange with ATLAS.

EUDICO was designed and is currently being implemented with the help of Java technology. It has a number of features that make it one of the most advanced annotation environments for dealing with multimedia language resources. It is closely coupled with the metadata browser, supports access to completely distributed resources, supports media streaming across networks, offers various synchronized viewers each representing the media and annotation data in stereotypic corpus-independent ways to the user, and allows to carry out complex searches on the data.

EUDICO was also designed to allow users working at different locations to collaborate on one resource. While one user could take care of the annotation of gestures, the other person could encode phonetics. Both should be able to immediately see what the other one is typing in at that moment. Full UNICODE support is essential for projects for documenting endangered languages, since the linguists are creating annotations with IPA and in languages such as Chinese, Arabic, Hindi, Cyrillic and others. This requires much care in designing the input methods.

A media-streaming component was realised which is based on Java Media Framework and which integrates the various streams (sound, video, texts) at the client. The software is built such that only the relevant media fragments will be sent across the network to reduce the load and costs.

4. Discussion

Browse and search tools that depend on standardized metadata descriptions to gain access to multiple distributed corpora will greatly enhance the possibility to quickly find required resources and compare between them. However it does demand an extremely flexible framework of tools for the annotation analysis itself since the annotation formats are disparate.

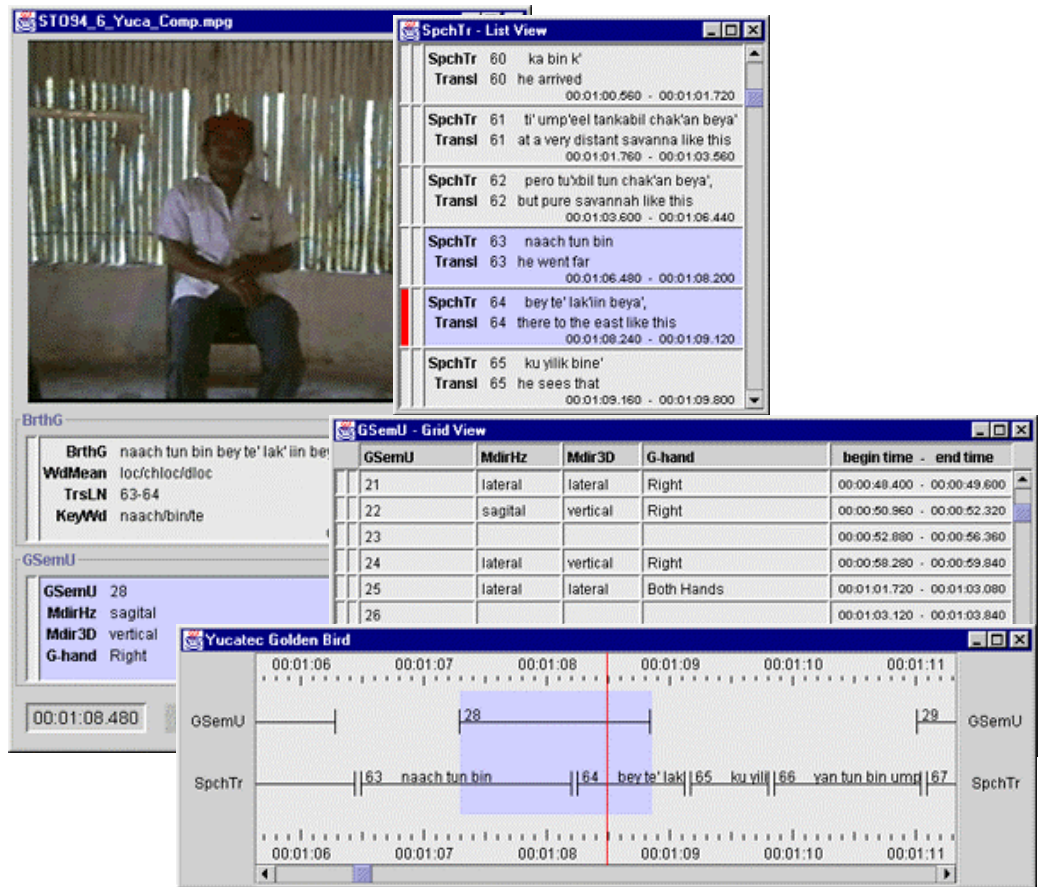


Figure 3 Several synchronised Eudico viewers

5. References

- [1] ISLE/IMDI <http://www.mpi.nl/ISLE>
- [2] P. Wittenburg, D. Broeder & B. Sloman: Meta-Descriptions for Language Resources - EAGLES/ISLE - A Proposal for a Meta-Description Standard for Language Resources. http://www.mpi.nl/ISLE/documents/papers/white_paper_1_1.pdf.
- [3] DOBES <http://www.mpi.nl/DOBES>
- [4] Corpus Gesproken Nederlands <http://www.now.nl/gw/introductie>
- [5] Text Encoding Initiative <http://www-tei.uic.edu/orgs/tei>
- [6] MPEG7 <http://www.cse.it/mpeg/standards.html>
- [7] EUDICO <http://www.mpi.nl/world/tg/lapp/eudico/eudico.html>
- [8] Shoebox <http://www.sil.org/computing/shoebox>
- [9] CHAT <http://childes.psy.cmu.edu/>
- [10] ATLAS <http://www.nist.gov/speech/atlas>

