

Advanced Web-based Language Archive Exploitation and Enrichment

Peter Berck, Albert Russel, Marc Kemps-Snijders, Peter Wittenburg

Max Planck Institute for Psycholinguistics
Wundtlaan 1
Nijmegen
The Netherlands

Peter.Berck@mpi.nl, Albert.Russel@mpi.nl, Marc.Kemps-Snijders@mpi.nl, Peter.Wittenburg@mpi.nl

Abstract

This article describes the linguistic archive set up at the MPI in Nijmegen. The MPI is in the progress of making it possible to exploit and enrich this archive over the internet with standard tools like a webbrowser. We describe how the different parts that make up the archive work together and the current state of the web based exploration. We also show the future direction for each of the different parts.

1. Introduction

At the MPI for Psycholinguistics a digital language archive has been set up during the last years that covers now about 40.000 sessions or resource bundles and other linguistic data types such as lexica, field notes and sketch grammars. Usually resource bundles combine different media files as video recordings, audio recordings, in some cases other signals such as eye movement recordings and several layers of (multimodal) annotations. Therefore, the language archive contains already far more than 100.000 individual language resources. The contributions to this archive come from different projects having very different objectives such as

- the DOBES programme founded by the VolkswagenFoundation where 24 documentation teams work all over the world to document languages that will become extinct in a few years time;
- the corpus of the ESF funded programme on adult language learners
- the Dutch National Spoken Corpus that covers annotated spoken utterances with about 10 Million words
- the corpora from the language and cognition department of the MPI covering gesture studies of various sort, interviews, recordings of elicited utterances and others in many different languages
- the corpora from the acquisition department of the MPI covering language acquisition corpora, longitudinal studies at different age groups and others also in many different languages.

Usually the included corpora are based on audio and video recordings, resources that take up much storage space so that the total archive already amounts to 11 TeraByte of data. Every year about 35 field trips are organized by MPI researchers and also the DOBES documentation teams are very active and are continuously integrating new data. Add to this that we will also open the archive for contributions from external projects that don't have the facilities themselves from summer 2005

on. We can assume that the amount of highly interesting language resources in the archive will grow very fast, i.e. an increase by a factor two within 3 years is estimated.

The whole archive is well organized according to the web-based IMDI catalogue principles (IMDI, 2005), i.e. the archive structure is organized with the help of IMDI corpus nodes in sub-corpora and the resource bundles, lexicons or other relevant linguistic objects are described with the help of the IMDI metadata descriptions¹. When either downloading the IMDI Browser or using normal WWW browsers and selecting the MPI corpus portal (CORPORA, 2005) one can browse and search in this MPI language archive domain without restriction since all metadata is open. The IMDI metadata descriptions include references to the resources (as is shown in figure 1) allowing suitable tools to be started from the browser. The structure of the corpora that are part of the archive are defined by the projects or individual researchers and of course they create the detailed metadata descriptions according to their insights.

Various checks that are carried out regularly guarantee a high degree of structural integrity of the archive. We transform different annotation and lexicon formats to one of a limited set of well-known formats such as EAF (ELAN, 2005), CHAT (CHILDES, 2005), Shoebox (SHOEBOX, 2005) and LMF (LMF, 2005) where possible and therefore achieve a high degree of format coherence in the archive. This all makes the MPI language resource archive an increasingly interesting source for linguistic analysis.

¹ It should be mentioned here that the IMDI domain of metadata described and registered language resources now covers contributions of more than 50 institutions amongst which are a number of well-known European language resource centers such as ELDA, BAS, Meertens, Lund University, Helsinki University, Florence University, ILC, ILSP, DFKI, various National Sign Language Communities etc. These contributions were created within the ECHO and the INTERA projects.

In this paper we will not discuss the mechanisms that we are applying to achieve long-term preservation of the

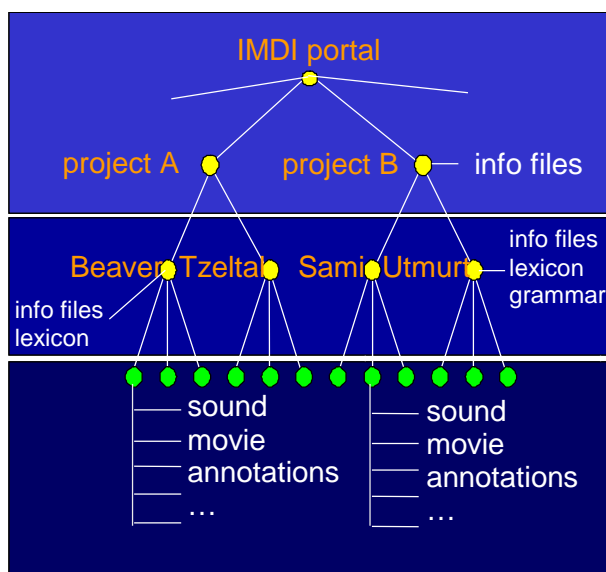


Figure 1 shows the IMDI archive/corpus organization principles. The yellow nodes are corpus nodes that can refer to information files of different sort. The green nodes represent detailed metadata descriptions of complex resources such as annotated media files or lexicons.

archival data. Nor will we report in detail on the management system that allows researchers to upload and modify metadata and resources in the archive. However, it is important to mention that we see a language resource archive in between traditional archives and modern digital libraries. While the primary concern of archives is the long-term preservation, digital libraries focus on access issues and dynamic content. Modern language resource archives have to take care of both aspects. Linguistic content has to be dynamic to a certain extent, since new linguistic insights for example may lead to modified and improved annotations and lexicon content. This notion was widely supported at a recent workshop of the DELAMAN network in which some of the major endangered languages and music archives are collaborating (DELANAN, 2005).

2. Archive Access Strategies

For all MPI resources different modes of access have to be provided. (1) It must be possible to browse and search in the metadata domain to find useful resources. The IMDI infrastructure allows to do so. It offers a number of options such as (a) browsing in the domain of linked IMDI/XML files either directly or facilitated by on the fly transformations to HTML; (b) carrying out structured search giving a high precision; (c) searching in full-text mode covering all metadata elements; (d) support for Google indexing; (e) support for services that are based on OAI metadata harvesting protocol (OAI, 2005). (2) Once found a suitable resource (bundle), it must be possible to download or directly play and visualize them. This is possible for audio, video and textual resources if the corresponding players are installed and also for

complex objects such as SMIL objects (SMIL, 2005) that are prepared offline showing videos with annotations as subtitles.

The described access modes refer to metadata and deal with individual resources or prefabricated objects such as SMIL which are treated by special players. Of course, access to the resources can only be given when access rights have been granted beforehand by the responsible researchers. The IMDI infrastructure comes along with an access management system that allows to define policies and set rights. However, an online digital archive has to offer more advanced web-based access methods that have the advantage that the user does not have to first download resources, organize them on the local machine and only then being able to continue the work. We distinguish between pure exploitation and enrichment tools that are based on server-side components. The latter will allow a direct manipulation of the content. However, archive content would be changed which is a critical operation. The only solution is to create first a copy in a temporary workspace where changes can be carried out and second upload the modified resource as a new version following specified rules and checks.

3. The Basket Principle

It is assumed that the user will make some form of pre-selection in the archive, since including all archive resources in a research activity does not seem probable (although it may be possible). The resources selected can come from various sub-corpora for example to carry out a comparative study or to determine cognate sets between languages. The selection may be done by using metadata browsing and/or searching. We call this selection the basket of resources that forms a virtual personal corpus for a certain researcher to carry out a certain research. We can assume that the linguistic encoding, i.e. the used tag and attribute sets and value ranges, are not harmonized. The selected resources may have annotation tiers that will contain the same type of information but have different tag labels such as “mo” or “morpho” for morphosyntactic information. Also the encoding within a tier can be different. In the most simple case we can assume values such as “n” and “no” both encoding the fact that the corresponding word is a “noun”. Both examples indicate the problems search engines will be faced with. Although within ISO TC37/SC4 we are working on a data category registry (LMF, 2005) that will contain many agreed linguistic concepts, we have to admit that it will take some time until the DCR will contain a critical mass of concept definitions, that it will take time until all tools will support referring to such concepts and that there will remain many differences between languages and linguistic theories. Therefore, we can assume that the problem of semantic interoperability will remain. Any tool that supports cross-corpus operations such as searching has to offer means to solve the lack in interoperability.

At the MPI we currently work on a number of web-based components that support working on such heterogeneous baskets: (1) ANNEX a component allowing to visualize

and play annotated multimedia recordings and structured texts and search in them; (2) LEXUS a flexible tool allowing to visualize arbitrarily structured lexica and to search in them; (3) COMEX a flexible tool allowing to draw and exploit semantic relations between arbitrary elements. For all components a style for a unified look and feel was recently worked out and is currently being implemented.

3. ANNEX Annotation Exploitation Tool

Multimodal annotation of multimedia files in general covers a number of time aligned media recordings such as video, audio and eye tracking and several layers of annotations. Since multimodal channels are seen as independent, any form of overlap in time between activities can occur. On the other hand linguistic tiers may exhibit hierarchal dependencies, i.e. they may inherit for example time boundaries from a parent tier. For a detailed analysis we refer to the papers of Bird & Liberman (Bird, 2001) and Brugman & Wittenburg (Brugman, 2001).

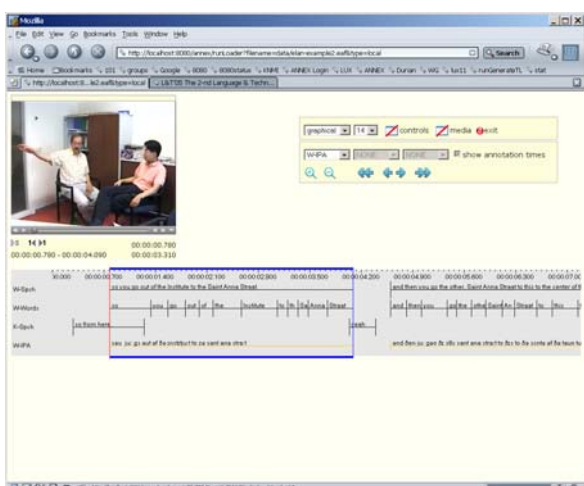


Figure 2 shows a screenshot of the current version of ANNEX. It offers different views on the annotations and all viewers are synchronized. In the figure a time line view is shown with a number of annotation tiers.

Crucial and new is the search engine that is provided with ANNEX, since we cannot assume a unified linguistic encoding according to an agreed coding handbook as described above. ANNEX assumes that the character encoding is done according to the UNICODE standard. It further assumes that the textual formats adhere to a few quasi standards such as EAF which is an XML format based on the Annotation Graph (Bird, 2001) and Abstract Corpus Model (Brugman, 2001) principles; CHAT which is the encoding format of the famous CHILDES initiative and SHOEBOX which is a tool widely used in field linguistics. ANNEX has filters that can parse these formats. To achieve interoperability at the linguistic level we have to give the user the possibility to quickly inspect all tiers and attributes that occur in the resources selected within the basket, i.e. to present the tier labels and the

values used in these tiers. It has to be indicated whether the linguistic concept represented by the label is taken from a registered data category registry since then detailed definitions will be available. A flexible mechanism is provided that allows users to map between tiers, i.e. create a persistent personal profile that contains “maps_to” relations that can be exploited by search engines. As was carried out in the ECHO project (ECHO, 2005) for metadata of different nature, these mappings have to be expanded in the index file to keep searching fast. Currently, we will not support other types of relationships such as “is_subclass_of” since we want first to better understand how ANNEX is being used in cross-corpus research and which level of granularity is required.

4. LEXUS Lexicon Tool

Due to the differences between languages, the purpose of a lexicon and differences in linguistic theories almost lexica exhibit a different structure and a large heterogeneity of lexical attributes (data categories) used. Therefore, ISO TC37/SC4 is defining a flexible lexicon model called LMF (Lexicon Markup Framework) (LMF, 2005). The definition work is in an advanced state and LEXUS is a first implementation of this emerging LMF standard. It allows users to flexibly create and visualize lexicon structures and content. LEXUS supports multimedia extensions, collaborative lexicon work and it can be used on notebooks as well as on servers as central installations. Also LEXUS supports filters for a few known lexical structures such as CHAT and SHOEBOX.



Figure 3 shows 3 screenshots of the LMF compliant LEXUS tool. It can visualize lexicon structures and contents in different forms as for example a tree. It further allows to visualize lexical content in different user defined styles.

It also allows to bootstrap a lexicon schema from a given XML lexicon file when no schema is provided.

LEXUS is already linked with data category registries such as the one from ISO, i.e. when creating a new lexicon or extending a lexicon one can re-use existing and well-defined data categories and in doing so increase the level of interoperability. Due to the many legacy lexica

that people want to use also LEXUS has methods to visualize lexical attributes, create mappings between them and make them persistent. In doing so it is possible to search across several lexica.

5. Future Work

We briefly described the first two components of a comprehensive web-based exploitation framework for IMDI based language resource archives. First versions of these components will become available in early 2005. The next steps are to add a component to present unstructured textual material such as sketch grammars and field notes in the same framework and to implement complex interaction options between ANNEX and LEXUS. A first simple interaction option was already implemented: the user can select a unit in an annotation and visualize the appropriate lexical entry. Finally, we will develop a component that allows users to comment on multimedia web-contents and to draw and exploit arbitrary relations between elements of visualized objects. Versioning aspects will be dealt with by the LAMUS system (Wittenburg, 2005) that is currently being developed. When LAMUS is full in use, it allows the users to make uploads and checks the validity of the operation, we will take steps to allow web-based manipulation of content. LEXUS was already designed to support web-based structure and content manipulation to support collaboration. With the help of the indicated tool set a new way of exploiting and working with language archives will become possible.

References

- Bird, St. Liberman, M. (2001): A formal framework for linguistic annotation. *Speech Communication* 33(1,2), pp 23-60, 2001.
- Brugman, H. Wittenburg, P. (2001), The application of annotation models for the construction of databases and tools. IRCS Workshop on Linguistic Databases, University of Pennsylvania. CHILDES (2005) <http://childes.psy.cmu.edu>
- CORPORA (2005) <http://corpus1.mpi.nl/BC/IMDI-corpora>
- DELAMAN (2005) <http://www.mpi.nl/delaman>
- ECHO (2005) <http://www.mpi.nl/echo>
- ELAN (2005) <http://www.mpi.nl/tools/elan.html>
- IMDI (2005) <http://www.mpi.nl/IMDI>
- LMF (2005) <http://www.tc37sc4.org/>
- OAI (2005) <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- SHOEBOX (2005) <http://www.sil.org/computing/shoebox>
- SMIL (2005) <http://www.w3.org/TR/REC-smil>
- Wittenburg, P., Broeder, D., Claus, A. (2005) LAMUS paper proposed to conference.