Language Archiving Technology at the MPI

Jacquelijn Ringersma, Paul Trilsbeek and Peter Wittenburg

Language Archiving Technology

Language Archiving Technology (LAT) is a software package meant to contribute to archive infrastructures. It focuses on open accessibility of language resources; it supports dynamic and continuously enriched collections according to the *Live* Archives ideas; it stresses the need for long-term archiving of our digital collections covering unique material about languages that will probably become extinct in a few decades and it follows the trend towards service oriented architectures.

LAT components consist of data management and ingestion tools (IMDI, LAMUS, AMS) and of archive enrichment and visualization tools (ELAN, ANNEX, LEXUS). The tools are developed and maintained by the MPI. All LAT products are or will become available under an Open Source license, and will be usable free-of-charge in academic research.

IMDI Infrastructure

The IMDI metadata set allows the structured description of language resources and to integrate these resources into an open, distributed metadata domain.

The IMDI infrastructure consists of an IMDI editor for the creation of metadata XML files, and IMDI browser supporting structured and unstructured search that can navigate IMDI based corpora and an IMDI web-application supporting browsing via normal web browsers, search and setting bookmarks to nodes and resources

LAMUS and AMS

The web based **Language Archive Management and Upload System** allows depositors to create and modify the logical structure of their corpus and to upload resources and metadata descriptions.

LAMUS ensures that the archive remains consistent and coherent. LAMUS checks the content of the IMDI files and performs a file type check. LAMUS attributes a unique resource identifier to every resource uploaded into the archive.

Users work in LAMUS workspaces in which they carry out their manipulations until a satisfying state is achieved. Uploading the workspace will lead to an update of the archive database, search indices for metadata and content and makes the data available for further utilization.

Metadata in the LR archive are open, however for the access to the resources users can define access rights to individuals or groups using the **Access Management System**. AMS is web based, access rights are defined on corpus nodes and specifications can be extended to those video, audio, image and textual resources that can be found under the selected node.

Language Resource Archive

The repository of the MPI contains different types of linguistic material: e.g. the DOBES endangered languages archive, the ESF second Learner corpus, the Dutch Spoken National Corpus, MPI's gesture corpora, MPI acquisition corpora and MPI language documentations of the language and cognition research group.

The archive covers more than 200.000 objects, mostly organized in sessions that are described with the IMDI-based metadata descriptions. Mostly, these sessions contain digitized audio/video signals and layers of annotations. In general the access to these resources is limited and can be made available upon request.

http://corpus1.mpi.nl

State of the Archive

• 23 TeraByte of Data

• over 200.000 objects of primary data (media), annotations, grammars, field notes, ethnographic notes

• DOBES archive of endangered languages 40.000 session and about 100.000 objects

• included formats: XML, HTML, Chat, Shoebox, PDF, Wav, Mpeg1, Mpeg2, Mpeg4

• A persistent Unique Resource Identifier for each archive object

Enrichment and Visualization

ELAN

ELAN is a professional tool, allowing the user to create and modify complex annotations on video, audio or textual resources. An annotation can be a sentence, word or gloss, a comment, translation or a description of any feature observed in the media. Annotations can be created on multiple layers, called tiers. Tiers can be hierarchically interconnected. An annotation can either be time-aligned to the media or it can refer to other existing annotations.

The textual content of annotations is always in Unicode and the transcription is stored in an XML format. ELAN is written in the Java programming language and the sources are available for non-commercial use.

ANNEX

ANNEX is a web-based tool that allows users to search in and visualize annotated media files (EAF, Shoebox/Toolbox and Chat). It offers synchronized viewers for video and audio streams and complex structured annotations. Several stereotypic viewers allow users to choose the most useful view on annotations. ANNEX supports video streaming; only selected fragments defined by annotations or time periods will be transmitted.

A powerful search engine has been developed which allows searching in the contents of the annotation files The results of the search can be visualized with ANNEX. The search engine provides everything from a simple text search over the annotations to complicated regular expressions over multiple annotations with multiple constraints. Special search modes for "space delimited tokens" and "%mor" tiers (from chat files) are also available. Looking for patterns between annotations is also supported by defining two or more patterns for annotations and the relation between these annotation on the same tier.

LEXUS

LEXUS is a flexible tool for lexical manipulations via the web or on a local system. LEXUS enables the creation of multi-media, encyclopedic lexica. Allowing to link audio, video and images to the lexical entries of the lexicon. LEXUS further facilitates the creation of semantic knowledge networks, through the relation functionality which allows to create relations between lexical entries or between attributes of the lexical entries.

LEXUS is compliant with the generic Lexical Markup framework (ISO TC37/SC4) and with the ISO DCR concept registry, thus enabling search through lexica, and merging of lexica.

New versions of LEXUS will allow collaborative workspaces, allowing different users to work on the same lexicon at different locations.

LAMUS **AMS** **IMDI infrastructure**

Archive Utility Layer

User Authentication | Access Management | Metadata Tools

Ontological Knowledge

The Archive

Domain of Registered Primary and Secondary Resources

Domain of Descriptive Metadata

Data Ingestion& Management

Primary Resources:
Texts
Images
Sound
Movies

Web-based Archive Utilization

Annotation Utilization | Lexicon Utilization | Text Utilization

User

Web-based Archive Enrichment

Media Annotation | Lexical Encoding | Web Commentary

● done
● in progress
● to start

ELAN for off-line work on annotations

ANNEX for visualizing annotated media files

LEXUS v1