SPLICR: A Web-Platform for Exploring, Querying and Distributing Linguistic Resources

Andreas Witt
(Institut für Deutsche Sprache, Mannheim)
Overview

- **Project Background: Aim of the Project**
- **Different Annotation Strategies**
- **Generalised Architecture for Sustainability of Linguistic Data (GENAU)**
- **Merging differently annotated corpora**
  - Transforming Single Rooted Trees
  - Transforming Annotation Graphs
  - Transforming Stand-off Annotations
- **The Platform SPLICR**
Principal Investigators:
Marga Reis, Erhard Hinrichs (Tübingen)

Project Members:
Christian Chiarcos (Potsdam), Timm Lehmberg (Hamburg)
Georg Rehm, Oliver Schonefeld, Andreas Witt (Tübingen)

Programming:
Johannes Dellert, Kilian Evang, Jonathan Khoo
Aim of the Project: Sustainability of Linguistic Data

- **Main goal:**
  - Resources acquired in long-term projects from three Collaborative Research Centres have to be converted in one format to be sustainably usable by researchers and applications

- **Additional aims:**
  - Provide unified access for the heterogeneous data acquired in the projects
  - General methodologies and ‘Rules of Best Practice’
The unusual starting point of the project

- In 2005 the German Science Foundation (DFG) initiated a joint project of three Collaborative Research Centres:
  - *Linguistic Data Structures*
    Research Centre 441 located in Tübingen
  - *Multilingualism*
    Research Centre 538 located in Hamburg
  - *Information Structure*
    Research Centre 632 located in Potsdam and Berlin
- Each Collaborative Research Center is a combination of ~15 research projects
Working areas / Areas of research

- Annotation formats
- Metadata
- Ontologies
- Corpus query
- Data visualization
- Legal Issues
Overview

• Project Background: Aim of the Project

• Different Annotation Strategies

• Generalised Architecture for Sustainability of Linguistic Data (GENAU)

• Merging differently annotated corpora
  • Transforming Single Rooted Trees
  • Transforming Annotation Graphs
  • Transforming Stand-off Annotations

• The Platform SPLICR

• Future developments
Resources and Annotation Schemes

- Researchers create linguistic data with a specific linguistic theory and a concrete research question in mind.
- This results in highly heterogeneous approaches to linguistic data handling and markup languages.
- All three research centres involved address this problem: at each site, a central project is assigned with the task of developing methods:
  - for the creation,
  - annotation and
  - analysis of linguistic data.
• Topic: Linguistic Data Structures
• Established: 1999
• At the moment ca. 100 researchers are involved
• projects investigating specific linguistic phenomena
  - with regard to general methodological issues, or
  - concerning a particular language or language family
• Almost all projects use corpora
• About 20 corpora have been created
• Despite the diversity of the corpora created in Tübingen, they all share the same generic data model

• All corpora are structured hierarchically

• A common annotation scheme, called TUSNELDA, was developed:
  - DTD
  - Annotation guidelines
Annotation Procedure

- Annotation: manual and automatic
- Embedded (inline) annotation, immediately modelling hierarchical structures by XML hierarchies
- Main Reason for using embedded markup:
  - Standard XML tools (such as XML editors and XML-parsers) are optimised for processing hierarchical XML structures with embedded annotation
TUSNELDA-Annotation: An Tibetan Example

khra·phru·gu  med·tshug |
child-Abs  NEG-exist

Translation: ‘Ø [=They] had no children.’

Structure of the annotation/Example annotation:

<s>
  <clause>
    <ntNode>...</ntNode>
    <tok id="v6">...</tok>
    <clauseCat>...</clauseCat>
  </clause>
  <punct> | </punct></s>
<s><clause><ntNode>
  <tok>
    <orth>khra•phru•gu</orth>
    <pos>NOM:anim~pers</pos>
  </tok>
  <ntNodeCat>NP</ntNodeCat>
  <desc><case>Abs</case></desc></ntNode>
<tok id="v6">
  <orth n="2">med-tshug</orth>
  <pos>VFIN</pos>
  <desc>... </desc>
  <feature type="part">Neg</feature>
  <frame>... </frame>
  <realframe>... </realframe>
  ...........</clause></s>
14 projects, all of them work empirically
- Written or transcribed spoken language
- Different linguistic research topics, e.g. language acquisition
  → The data differ with respect to many dimensions
- To provide a unified access the EXMaRALDA system was developed
EXMaRALDA’s Basic Data Model

- Based on the “Single Timeline, Multiple Tiers” model (annotation graphs framework)
- Individual descriptions (events) are grouped into a number of tiers (or layers)
- Ideally, a start and an end point of each event is marked on a single, fully ordered timeline
- This model is also used in other systems and tools, e.g. Praat, ELAN, TASX
- Directed, acyclic graphs
- XML as storage format, no hierarchies
- Ontologically empty framework, i.e. it abstracts from linguistic theories
Example

Transcription: [v]
Description: [nv]
Annotation: [sup],[en],[pho]

| DS [sup] | faster    |
| DS [nv]  | right hand raised        |
| FB [v]   | Alors ça dépend ((cough)) un petit peu. |
| FB [en]  | That depends then, a little bit. |
| FB [pho] | [étipô:]                |
Information Structure concerns the means used by the speaker or writer to structure discourse and utterances.

Languages differ a lot with regard to the means to express Information Structure, e.g., intonation, word order, etc.

Empirical base: Different types of corpora, languages, and annotations.
Potsdam’s Corpus Interchange Format

- Stand-off XML-annotation
  - allows for conflicting hierarchies
  - extensive use of XLinks, XPointers to link files
- Generic XML elements
  - based on upcoming ISO-Standard Linguistic Annotation Framework
  - flexible enough to encode all kinds of data structures
  - flat lists of XML elements; hierarchies encoded by XLinks

  - `<mark>` markables
  - `<struct>` hierarchical structures
  - `<feat>` annotations
Overview

- Project Background: Aim of the Project
- Different Annotation Strategies
- **Generalised Architecture for Sustainability of Linguistic Data (GENAU)**
  - Merging differently annotated corpora
    - Transforming Single Rooted Trees
    - Transforming Annotation Graphs
    - Transforming Stand-off Annotations
- The Platform SPLICR
Motivation: Why yet another format?

- TUSNELDA, EXMARaLDA and PAULA all generalise over project-specific data models and formats
- We need: a data model which generalises over TUSNELDA, EXMARaLDA and PAULA.
- This model must be applicable for all the language data already annotated:
  - Hierarchical annotations with embedded markup
  - Graph based annotations
  - Distributed markup using stand-off techniques
- Exchange format should be as simple as possible
Generalised Architecture for Sustainability of Linguistic Data

• “Genau”:
  • Generalisierte Nachhaltigkeitsarchitektur für linguistische Daten
    (Generalised Architecture for Sustainability for linguistic data)
  • German for: accurate, close, correct, definite, demanding, detailed, exact, faithful, fine, just, minute, pedantic, precise, right, strict

• Provides a format for an Unified Linguistic Annotation
• The original annotation format becomes irrelevant
Genau: Format description

- In general, the Genau-Format can be modeled by means of Multi-rooted trees (MRTs)
- MRTs are neither as constrained as a tree, nor as open as an unrestricted graph
- Storing:
  - in an XML-Database (for processing, e.g. querying)
  - in individual XML files (as a sustainable interchange format)
- Each file represents all the information related to a single linguistic annotation layer
Transformatios into the Genau-Format

- Corpora annotated based on the hierarchical model are analysed semi-automatically.
- After the analyses, information on the layers is included in the (still single rooted) XML document instance.
- In the next step, the hierarchically annotated corpora is split into individual XML files.
- Timeline-based corpora are split using another tool in order to separate the graph annotations.
- Standoff annotations and the text are merged.
Overview

- Project Background: Aim of the Project
- Different Annotation Strategies
- Generalised Architecture for Sustainability of Linguistic Data (GENAU)
- Merging differently annotated corpora
  - Transforming Single Rooted Trees (small tool one)
  - Transforming Annotation Graphs
  - Transforming Stand-off Annotations
- The Platform SPLICR
A tree-based XML document (such as a corpus) with multiple annotation layers may need to be separated.

“Leveler” is a pipeline for XML-document transformations.

Leveler serves two purposes:

1. Moving PCDATA text to attributes (thereby separating any PCDATA annotations from the actual primary data).
2. Splitting the corpus into different files according to the different layers of annotation (e.g., syntactic, morphological, etc.).

The transformations are carried out using XSLT.

The configuration files for XSL-processing are created in a web application.
Examples for Text Transformations

- Original
  
  `<w>Peter</w>`
  `<pos>NN</pos>`
  `<punct>!</punct>`

- “Real PCDATA”
  
  `<w>Peter</w> (identity)`

- “Annotation”
  
  `<pos leveler:text="NN"/>`

- Mixed
  
  `<punct leveler:text="!"/>!`
Example: Text Transformation

<tok id="s_18_n_2">
<orth>Landesvorsitzende</orth>
<pos func="HD">NN</pos>
</tok>

<tok id="s_18_n_2">
<orth>Landesvorsitzende</orth>
<pos func="HD" leveler:text="NN" />
</tok>
Leveler Pipeline

Step 1
Text Transformation

Original Corpus XML

Text Transformation (XSLT)

Web Step 1

Text Transformation Configuration File

Step 2
Layer Specification

Text-Transformed Corpus

Layer Specification (XSLT)

Web Step 2

Layer Specification Configuration File

Layer 1
Layer n

Step 3
Splitting

Layer-Specified, Text-Transformed Corpus

Layer Splitting (XSLT)

Layer 1
Layer n
Overview

- Project Background: Aim of the Project
- Different Annotation Strategies
- Generalised Architecture for Sustainability of Linguistic Data (GENAU)
- Merging differently annotated corpora
  - Transforming Single Rooted Trees
  - Transforming Annotation Graphs (small tool two)
  - Transforming Stand-off Annotations
- The Platform SPLICR
<?xml version="1.0" encoding="UTF-8"?>

<basic-transcription>
  <head> <!-- ... --> </head>
  <basic-body>
    <common-timeline>
      <ti id="T0" time="0.0"/> <ti id="T1"/> <ti id="T2"/> <ti id="T3"/> <ti id="T4"/> <ti id="T5"/>
    </common-timeline>
    <tier id="TIE0" speaker="DS" category="sup" type="a">
      <event start="T1" end="T3"> faster </event>
    </tier>
    <tier id="TIE4" speaker="FB" category="V" type="t">
      <event start="T2" end="T3"> Alors ça </event>
      <event start="T3" end="T4"> dépend ((cough)) </event>
      <event start="T4" end="T5"> un petit peu. </event>
    </tier>
    <tier id="TIE5" speaker="FB" category="en" type="a">
      <event start="T3" end="T5"> That depends, then, a little bit. </event>
    </tier>
    <tier id="TIE6" speaker="FB" category="pho" type="a">
      <event start="T4" end="T5"> [ɛliːpø:] </event>
    </tier>
  </basic-body>
</basic-transcription>
The relations of segments and annotation are conveyed by attribute references to time line items.

**timeline**

annotation (speaker 1)

**segments (speaker 1)**

annotation (speaker 1)

annotation (speaker 2)

**segments (speaker 2)**

annotation (speaker 2)

annotation (speaker 2)
EXMARaLDA
*segmented transcription*

time-based format

GENAU

hierarchy-based format
Okay.

faster

Very good, very good.

That depends, then, a little bit.

[\textipa{\epsilon ip\textipa{\textordmasculine o}:}]

Okay.

Très bien, très bien.

Alors ça
dépends ((cough))

un petit peu
Splitter Usage

XSLT 2.0 Processor

transcription_s.xml

splitter.xsl

main document

transcription_s_split.xml

directory containing the multi-rooted tree

SPLICR: A Web-Platform for Exploring, Querying and Distributing Linguistic Resources
Overview

- Project Background: Aim of the Project
- Different Annotation Strategies
- Generalised Architecture for Sustainability of Linguistic Data (GENAU)
- Merging differently annotated corpora
  - Transforming Single Rooted Trees
  - Transforming Annotation Graphs
  - Transforming Stand-off Annotations (small tool three)
- The Platform SPLICR
Transformation approach

- Convert as much of the hierarchies as possible to a leaf-ordered multi-rooted tree.
Transformation approach

- Convert as much of the hierarchies as possible to a leaf-ordered multi-rooted tree.
- Where this is not possible due to crossing branches, encode relations using XLink
- Use XSLT
Die einstige Fußball-Weltmacht

The GENAU data model: A multi-rooted tree (MRT)
Example Data in the Format “Genau”

```
<genau>
  <pos value="ART"><tok id="t5">Die</tok></pos>
  <pos value="ADJA">
    <tok id="t6">einstige</tok></pos>
  <pos value="NN">
    <tok id="t7">Fußball-Weltmacht</tok></pos>
</genau>

pos.tok.xml

<genau>
  <cat value="NP">
    <const id="s2">
      <tok id="t5">Die</tok>
      <tok id="t6">einstige</tok>
      <tok id="t7">Fußball-Weltmacht</tok>
    </const>
  </cat>
</genau>

cat.const.tok.xml
```
Overview

- Project Background: Aim of the Project
- Different Annotation Strategies
- Generalised Architecture for Sustainability of Linguistic Data (GENAU)
- Merging differently annotated corpora
  - Transforming Single Rooted Trees
  - Transforming Annotation Graphs
  - Transforming Stand-off Annotations
- The Platform SPLICR
SPLICR: An overview

- **SPLICR: Sustainability Platform for Linguistic Corpora and Resources**

- The sustainability platform consists of a front-end and a back-end

- **The front-end**
  - is the user visible part and is realized using Java Server Pages (JSP), JavaScript and Ajax technologies
  - runs in the user's browser
  - provides functions for searching and exploring metadata records and corpus data

- **The back-end is a web application that runs on top of the Tomcat application server**

- In addition, a staging area (the data repository) contains the normalized corpora
The staging area

- All corpora and metadata are converted in a normalized form
- A manifest file is generated for each corpus
- The manifest is a simple XML format and acts as a corpus inventory list
- Manifest files are generated semi-automatically
- Each corpus consists of five parts:
  - the manifest file,
  - multiple files that contain the processed corpus data,
  - multiple files that contain the metadata record
  - the original and unchanged corpus files, and
  - stylesheets, logfiles etc.
Resources in the staging area

Staging area (directory tree on a server) contains:
- Corpus data (separated into individual annotation layers)
- eTEI metadata records
- Original corpus data (original files, audio data etc.)
- Manifest files (XML-based inventory list of a corpus)

SFB 441 (University of Tübingen)
- Corpus data (normalised)
- Manifest file (XML-based)
- Metadata records in eTEI
- Custom normalisation tools
- eTEI editor
- Original corpus data (arbitrary XML-based formats)
- Original manifest

SFB 538 (University of Hamburg)
- Original corpus data (arbitrary XML-based formats)
- Original metadata

SFB 632 (University of Potsdam)
- Corpus data (normalised)
- Manifest file (XML-based)
- Metadata records in eTEI
- Custom normalisation tools
- eTEI editor
- Original corpus data (arbitrary XML-based formats)
- Original manifest

Corpus Processing and Normalisation (at different sites)
The directory structure of the staging area (excerpt)

```
/corpora/SustainabilityPlatform/LanguageResources
  name-mapping.xml ................. contains mapping of directory names to display names
  UniversityOfHamburg
    name-mapping.xml
    SFB538 .... subtree of the staging area maintained by SFB 538 (University of Hamburg)
      ...
  UniversityOfPotsdam
    name-mapping.xml
    SFB632 .... subtree of the staging area maintained by SFB 632 (University of Potsdam)
      ...
  UniversityOfTübingen
    name-mapping.xml
    SFB441 .... subtree of the staging area maintained by SFB 441 (University of Tübingen)
      name-mapping.xml
      A01 ............... directory for corpora created by the SFB 441 project A01
        TüPP-DZ
          ...
        TüBA-DZ
          2003-08-31T16-47+01-00_R1 ............... different versions of a corpus
            ...
          2005-08-31T16-47+01-00_R2
            manifest.xml ...................... contents of a corpus directory
            Corpus-* ..................... corpus data processed by our sustainability project
            Metadata-* ................ metadata processed by our sustainability project
            Original .................... the original files associated to a corpus
              ...
            Transformation ................. transformation stylesheets and logfiles
```
SPLICR: A Web-Platform for Exploring, Querying and Distributing Linguistic Resources

The web-based Platform: Front-end and back-end

Front-end/Browser

Dispatcher (JSP)

File vault
(for additional files)

OntoClient
and ontologies

User
data

Resource
data

Access
restrictions

SQL system database

XML database

eTEI-based
metadata

Corpus
data

AnnoLab

Front-end/Browser administration interface

HTTP/REST

Importer service

Importer traverses staging area

Staging area (directory tree on a server) contains:
- Corpus data (separated into individual annotation layers)
- eTEI metadata records
- Original corpus data (original files, audio data etc.)
- Manifest files (XML-based inventory list of a corpus)
The Back-end

- The back-end hosts the JSP files and related data.
- It accesses two different databases,
  - the corpus database and
  - the system database.
- Furthermore, all additional files (e.g. original corpus data files, documentation, transformation scripts) are stored in the file system.
- Several servlets provide means for exchanging information between the front-end and the back-end.
- The back-end is implemented as a web application that runs on top of Apache's Tomcat servlet container.
- The corpus database is an eXist XML database, extended by the AnnoLab system (Richard Eckart and ElkeTeich)
- A System database
System database

- Uses a relational database (MySQL)
- Contains data about user accounts and acts as a catalogue for corpus data
- Stores information about
  - single files in a corpus,
  - resource groups (i.e., corpora) and
  - access rights.
- A specific user can only access a specific resource if the permissions for this user/resource pair allow this operation
Carrying out a query

- The front-end sends a JSON representation of the query and a list of the corpora currently selected by the user to the query dispatcher servlet.
- A servlet transforms the query into XQuery by generating, for every single file of all selected corpora, a dedicated XQuery expression.
- This set of XQuery expressions is linked to a query job which is executed using a worker-thread of the query service component.
- At the same time, a unique query ID is returned to the front-end, which will start polling results.
- The XQuery expressions are run sequentially against the corpus database.
- Results are buffered within the back-end until the front-end fetches them.
Advantages of this query process implementation

- The user can already start exploring the first result even though the system is still running queries on the remaining files.
- A query monitor exists in the administration area of the front-end.
- It allows the administrator to display all currently running query jobs with additional details such as average query runtime per file and estimated remaining total runtime.
- The administrator can cancel query jobs.
The Front-end supports answering the questions

- Which linguistic resources are stored in the platform?
- Can one or more of these corpora be used as empirical data bases for a specific research question one is working on?
- What is the extent of the annotations of these resources and do they cover what is needed for one's research endeavor?
A typical usage scenario

Linguist has a specific research question

Login

SPLICR

Exploration of the metadata of all corpora

Exploration of a specific corpus (using query options)

Download of a corpus

Logout

Linguist can work with corpus offline
Front end functionalities

- As soon as one or more corpora are selected the user can access the query interface which is based on two main concepts.
- Three different kinds of search widgets
  - full-text search,
  - concept search, and
  - tree fragment search.
- The query interface supports multiple tabs that can be added and deleted at will.
Metadata Exploration

- As soon as a user logs onto the system a complete list of resources currently stored in SPLICR is presented.
- Drop-down menus can be used to filter the list.
- A click on the “information” icon expands the row that contains the name of the resource and its affiliation.
- This expanded view shows:
  - a brief description of the corpus,
  - its version,
  - the annotation layers, and
  - The number of files associated with this resource.
Listing of available resources – Selecting resources

<table>
<thead>
<tr>
<th>Resource Type</th>
<th>Organisation</th>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dik</td>
<td>Universität Hamburg, SFB 538</td>
<td>Project K02: Interpreting in Hospitals</td>
</tr>
<tr>
<td>SkanSem</td>
<td>Universität Hamburg, SFB 538</td>
<td>Project K05: Semi-Communication and Receptive Multilingualism in Scandinavia</td>
</tr>
<tr>
<td>KonSim</td>
<td>Universität Hamburg, SFB 538</td>
<td>Project K06: Coherence in Interpreter-Mediated Discourse</td>
</tr>
<tr>
<td>Potsdam Commentary Corpus, public subsection</td>
<td>Universität Potsdam, SFB 832: Information Structure</td>
<td>Project D01: Linguistic Database for Information Structurally Annotated Corpora</td>
</tr>
<tr>
<td>Tatian corpus</td>
<td>Universität Potsdam, SFB 832: Information Structure</td>
<td>Project B04: The role of information structure in the development of word order regularities in German</td>
</tr>
<tr>
<td>TüPoDis-Korpus</td>
<td>Universität Tübingen, SFB 441: Linguistic Data Structures</td>
<td>Project B09: Local and Temporal Diakits in the Romance Languages: History and Variation</td>
</tr>
</tbody>
</table>

Audio recordings and annotated transcriptions of semi-communication and receptive multilingualism in present-day Scandinavian languages (Danish, Norwegian, Swedish). The audio recordings have been collected at several classes and from radio transmissions. They have been transcribed and annotated according to the HINT-standard.

Version: 12th June, 2008
Annotation layers: Contribution of Events, Contribution of Utterances, Comments 1, Events, English Translation, Code Switching, German Translation, Phonological Comments

Number of associated resource files: 468
<table>
<thead>
<tr>
<th>Resource Type</th>
<th>Organisation</th>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiK</td>
<td>Universit&quot;at Hamburg, SFB 538</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Project KG2: Interpreting in Hospitals</td>
<td></td>
</tr>
<tr>
<td>SkanSem</td>
<td>Universit&quot;at Hamburg, SFB 538</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Project KC5: Semi-Communication and Receptive Multilingualism in Scandinavia</td>
<td></td>
</tr>
</tbody>
</table>

Audio recordings and annotated transcriptions of semi-communication and receptive multilingualism in present-day Scandinavian languages (Danish, Norwegian, Swedish). The audio recordings have been collected at several classes and from radio transmissions. They have been transcribed and annotated according to the HIAT-standard.

Version: 12th June, 2008
Annotation layers: Contribution of Events, Contribution of Utterances, Comments I, Events, English Translation, Comments II, German Translation, Phonological Comments
Number of associated resource files: 468

<table>
<thead>
<tr>
<th>Resource Type</th>
<th>Organisation</th>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>KonSim</td>
<td>Universit&quot;at Hamburg, SFB 538</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Project K06: Coherence in Interpreter-Mediated Discourse</td>
<td></td>
</tr>
<tr>
<td>Potsdam Commentary Corpus, public subsection</td>
<td>Universit&quot;at Potsdam, SFB 632: Information Structure</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Project D01: Linguistic Database for Information Structurally Annotated Corpora</td>
<td></td>
</tr>
<tr>
<td>Tatian corpus</td>
<td>Universit&quot;at Potsdam, SFB 632: Information Structure</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Project BC4: The role of information structure in the development of word order regularities</td>
<td></td>
</tr>
<tr>
<td>T&quot;uPoDia-Korpus</td>
<td>Universit&quot;at T&quot;ubingen, SFB 441: Linguistic Data Structures</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Project BC9: Local and Temporal Deixis in the Romance Languages: History and Variation</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resource Type</th>
<th>Organisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Super Corpus</td>
<td>Universit&quot;at T&quot;ubingen, SFB 441: Linguistic Data Structures</td>
</tr>
</tbody>
</table>
Metadata Exploration

- If the user wants to know more, the hyperlink “contents and download” switches to a view that lists all files that belong to a corpus.
Contents and download
## Linguistic Corpora and Resources

**University of Tübingen, SFB 441: Linguistic Data Structures**

**Project B08: Corpus-based Analysis of Local and Temporal Deictics in (Spontaneously) Spoken and (Reflected) Written Language**

### Processed Data
- Size: 2001948 bytes
- Number of files: 38

### Annotation Layer: Editorial Notes
- Size: 583444 bytes
- Number of files: 12

### Annotation Layer: Deictics
- Size: 623679 bytes
- Number of files: 13

1. Corpus-Deictics-Dataset_01_of_13-File_1_of_1.xml
2. Corpus-Deictics-Dataset_02_of_13-File_1_of_1.xml
3. Corpus-Deictics-Dataset_03_of_13-File_1_of_1.xml
4. Corpus-Deictics-Dataset_04_of_13-File_1_of_1.xml
5. Corpus-Deictics-Dataset_05_of_13-File_1_of_1.xml

### Annotation Layer: Conversation
- Size: 794825 bytes
- Number of files: 13

### Metadata
- Size: 20157 bytes
- Number of files: 6

- Metadata-Annotation-Conversation-File_1_of_1.xml
- Metadata-Annotation-Deictics-File_1_of_1.xml
- Metadata-Annotation-EditorialNotes-File_1_of_1.xml
- Metadata-Corpus-File_1_of_1.xml
- Metadata-PrimaryData-File_1_of_1.xml

### Original Data
- Size: 1339494 bytes
- Number of files: 44

- b8-interviewheader-metadaten.xml
- b8-interviews-BG1-metadaten.xml
- b8-interviews-BG2-metadaten.xml
- b8-interviews-BH-metadaten.xml
- b8-interviews-BJ-metadaten.xml

### Transformation Data
- Size: 52082 bytes
- Number of files: 26
Multiple Methods of Querying Corpora

- The target users (i.e., linguists) are (in general) not proficient in XML query languages such as XPath and XQuery,

- Therefore, intuitive query interfaces that generalize from the underlying data structures and querying methods are presented:
  - Full-Text Search: The full-text search query widget can be used to find certain words or simple patterns in corpora. Matches are highlighted in the result browser.
  - Concept Search: The concept search query widget presents a list of linguistic concepts that are contained in the individual annotation layers that make up a corpus.
  - Tree-Fragment Search: an interactive editor for constructing linguistic tree fragments that can be queried against the currently selected corpus
Example 1: The concept-search
Example 2: The tree-editor for constructing queries
Presentation of the search results

- We provide three different query widgets that can be used to search and query corpora.
- The results of these queries are visualized by the result browser that offers four different display modes:
  - plain text view
  - XML view
  - box view
  - tree view
Results view: tree view
Results view: box view
Summery:

- Project Background: Aim of the Project
- Different Annotation Strategies
- Generalised Architecture for Sustainability of Linguistic Data (GENAU)
- Merging differently annotated corpora
  - Transforming Single Rooted Trees
  - Transforming Annotation Graphs
  - Transforming Stand-off Annotations
- The Platform SPLICR
SPLICR: A Web-Platform for Exploring, Querying and Distributing Linguistic Resources