CLARIN-D Report R 8.2
Documentation of AP8 preparation activities

Marilisa Amoia, Hannah Kermes, Jörg Knappen, José Manuel Martínez Martínez, Elke Teich, Mihaela Vela

May 2013
1 Introduction

CLARIN-D as an infrastructure for language resources and technology has an explicit user oriented mission. The fundamental challenge is to open new potential users for the possibilities such an infrastructure can offer them for their research. Prejudices and obstacles with respect to language resources and technology have to be addressed and the potential user has to be made acquainted with their use and usage.

During the first year of the project phase we identified three different contexts with respect to training and eduction:

- integration of language resources and language tools into university degree programs,
- training of teaching staff,
- qualification of researchers.

In the framework of these three contexts, we worked on the following measures:

- a Teaching and Learning Material Collection (TeLeMaCo)—a collaborative portal for all kinds of training and teaching materials relevant to linguistics and digital humanities,
- individual hostings for researchers,
- training modules and tutorials,
- cooperation with AP7 (Support & Helpdesk) and DARIAH.

In the following we report on activities with regard to these measures.

The appendix contains reports of the hostings, programs of the PhD days and the workshop of AP7 and AP8.
2 Teaching and Learning Material Collection (TeLeMaCo)

Nowadays large amounts of data, resources and tools are available on the web for many specialized domains that are primarily accessible to researchers. However, since academic resources are distributed over several sites that are not necessarily interlinked and are often not represented in a common standard, gathering relevant resources still remains a cumbersome task for academic users.

As the range of academic materials available on the internet is steadily growing, there is an increasing awareness that discipline-specific “social networks” should be created that support the exchange of scholarly materials. Here, some disciplines are more advanced than others in setting up such support. For instance, while the field of computational linguistics is fairly well equipped with platforms such as the ACL web registry\textsuperscript{1} or the ELRA catalog\textsuperscript{2} and many others [Lemnitzer et al.(2009)], the broader field of linguistics cannot draw upon similar sources of structured materials.

In recent years, there have been selected efforts to build digital repositories for linguistic purposes, such as The Language Archive (TLA) [Drude et al.(2012)] or the Virtual Language Observatory (VLO) [Uytvanck et al.(2012)]. Both are platforms allowing to conveniently manage both collection and retrieval of large amount of linguistic resources and tools based on a common standard for metadata representation. For linguistic processing, there is, e.g., the WebLicht platform [Hinrichs et al.(2010)] that further provides facilities for text and speech processing and supports customization of processing pipelines ranging from annotation to parsing and semantic analysis. However, while it becomes much easier to find language resources (e.g., corpora) and tools (e.g., taggers, parsers) and to get some support in processing (as exemplified by WebLicht), it remains relatively difficult to find relevant training and learning materials. While such materials are potentially available on the web, there are only a few convenient portals that provide a centralized access to them.

TeLeMaCo aims at a centralized repository and portal for learning and teaching materials dealing with the application of language resources and tools to be used by linguists. In the following, we describe the basic functionalities of our system as well as the schema we employ for annotating training and teaching materials. Finally, we discuss the use of our system, focusing on search/query.

2.1 TeLeMaCo

Our portal (Teaching and Learning Material Collection)—TeLeMaCo—provides access to the following types of material related to language resources and tools:

- technical documentation (including tutorials, quick starts, etc.),
- learning material for self study,
- short teaching modules (2–4 hours) that can be integrated in existing courses,

\textsuperscript{1}http://aclweb.org/aclwiki/index.php?title=Resources
\textsuperscript{2}http://universal.elra.info/
full courses covering a broader spectrum of language resources and tools or focusing on specific topics of application of language resources and tools.

TeLeMaCo is essentially designed as a “social network”. Registered users may both contribute descriptions of learning and teaching material they make available on the web and/or use TeLeMaCo to find relevant material as well as provide comments on the material referenced by TeLeMaCo.

The interaction model behind TeLeMaCo distinguishes between three types of users with different permissions:

- **Administrator**: the technical staff responsible for system administration.
- **Author**: the owner of learning material published in the hub. Users of this type are responsible for the semantic annotation associated with the resources.
- **User**: the customer querying the database. Users of this type can further rate learning content in the hub by writing a comment or answering to a questionnaire.

### 2.2 Annotation of training and teaching materials

The prerequisite for successful search in a data repository is a meaningful annotation, which is true of course also for the types of content that we are concerned with here (cf., e.g., [Bärenfänger et al.(2011)]).

![Figure 2.1: Content Annotation in TeLeMaCo](image_url)
The categories of annotation used in TeLeMaCo are motivated from a teaching and learning perspective and include:

- standard categories like author, title, publication year, and URL;
- categories capturing pedagogically relevant features such as prerequisites, target audience, objectives, and level;
- bibliography facilities, e.g., bibtex\(^3\) and bibtex type;
- topic-relevant features, such as description;
- technical features, e.g., language and media type of the resource;
- accessibility information such as license and access.

The values of these features are directly assigned by the author of the learning content. Figure 2.1 shows a sample annotation.

A resource may be associated with user/learner feedback that directly addresses these categories, in particular the ones relevant from learning or pedagogical perspectives. A user can thus express her/his opinion about a resource by answering questions like What level has the resource in your opinion? or Did you reach the objectives given? Furthermore, each resource is associated with a list of keywords.

The annotation scheme of TeLeMaCo will be linked to the Component Metadata Framework (CMDI) \([\text{Broeder et al.}(2011), \text{Durco et al.}(2012)]\) that is being implemented within the CLARIN\(^4\) community. This will allow an easy integration of our system into linguistic repositories such as the VLO.

### 2.3 Querying

The database of TeLeMaCo can be queried through the web interface shown in Figure 2.2. It allows for fine-grained queries. So for instance, the user can retrieve learning material by setting search criteria such as author name, date, version, but also topic, difficulty level of the learning content or preferred media and whether the resource is open or requires a license.

In the future, we plan to automatically extract additional topic-relevant keywords from the resource description to be able to expand queries with synonyms and semantically related concepts and thus improve search results.

The system is still in the development phase and is not yet freely accessible on the web. However, we have started a test phase involving several academic institutions.

---

\(^3\)Our system automatically generates the bibtex entry for a resource on the basis of the information encoded in the standard category slots.

\(^4\)http://clarin.eu
Advanced search

Title: 
Author: 
Keywords: Phyton
Language: german
Objective: 
Audience: Linguistics
Prerequisites: 
Level: beginner
Type: 
Media: video
URL: 
Licence: 
Access: open academic
Description: 
Search

Figure 2.2: TeLeMaCo Search Interface.
3 Hostings

Hostings are a measure to enable researchers from the humanities with a dedicated research interest to learn from the expertise of one of the CLARIN-D centers while visiting for a longer stay (up to two months). It provides financial aid for a research stay at one of the CLARIN-D centers. Requirements for the funding are

- a concrete research plan
- a report about the stay
- a presentation of the work at the hosting center
- the integration of resulting resources or tools in CLARIN-D (if applicable)

During the reporting period, two hostings were financed by CLARIN-D:

- Oleg Kapanadze (Faculty of Humanities, Department of German Language and Literature, Tbilisi State University, Georgia)
  - research stay: April 19 to June 19, 2012
  - hosting center: Universität des Saarlandes
  - research: development of a parallel treebank Georgian, Russian, Ukrainian, German (GRUG) ([http://fedora.clarin-d.uni-saarland.de/grug/index.html](http://fedora.clarin-d.uni-saarland.de/grug/index.html)) (the report on the research in the appendix [6])

- Marco Passarotti (CIRCSE Forschungszentrum, Università Cattolica del Sacro Cuore)
  - research stay: February 1-28, 2013
  - hosting center: Universität Tübingen
  - research: integration of the Index Thomisticus Baumbank into CLARIN-D ([http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/index-thomisticus-baumbank.html](http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/index-thomisticus-baumbank.html))
4 Training modules and tutorials

4.1 CLARIN-D Doktorandentage (PhD days)

The CLARIN-D PhD days are intended as a small series of training events. The target group are young researches (mostly PhD students). Each of the single events has a specific thematic focus (e.g., corpora; speech; historical data).

The first CLARIN-D PhD days *Corpora* took place on March 25–26, 2013 at the IMS, University of Stuttgart. The event consisted of three small blocks focusing on basic corpus linguistic methods including corpus search with regular expressions and pattern-matching within the framework of CQP-web, syntactic annotation and corpus search using WebLicht and other CLARIN-D resources, and the statistical analysis using R. The program of the workshop can be found in the appendix.

All participants had to send in an abstract prior to the workshop stating their research interest, prior knowledge and expectations. This information was used to adapt the content of the workshop to the specific needs of the participants and to include examples from their research area wherever possible. The scientific background of the participants of this first event was mainly linguistics and literature (including socio-linguistics, typology, syntax, semantics, translation studies, diachronic studies, computational linguistics ...). Languages included German, Romance languages and Slavic languages. Prior knowledge about language resources and language technology ranged from beginner to advanced.

The overall feedback of the participants was positive. An important issue was the possibility to query annotated corpora in a sophisticated way and to be able to sort and group results to obtain relevant frequencies and frequency distributions.

There was also great interest in annotating their own text corpora and in the possibility to host them at CLARIN-D centers. As a first concrete result, one of the participants will use the presented CQPweb platform for a course on corpus linguistics at her home institute in April 2013.

Follow-up events with other thematic foci are planned for the next project year, e.g., in the area of speech data and historical texts.

4.2 Tutorials and Courses

**Tutorial: Tutorial for the content management system of the CLARIN-D webpage (Joomla)**

Jens Stegmann
CLARIN-D-Entwicklertreffen, IDS Mannheim, 13/06/2012


- Use of Backend
- Structure of the webpage
- Relaunch etc.
Tutorial: Korpusbasierte Sprachreflexion mit Online-Ressourcen – Eine praktische Einführung
Heike Zinsmeister
Workshop: Korpusbasierte Sprachreflexion, Paris, 16/06/2012
http://didatic.net/article.php3?id_article=581
- Linguistic online queries on DWDS and IDS corpora
- Annotation in WebLicht
- Queries in TigerSearch

Tutorial: Syntactic annotation and corpus search with CLARIN-D resources
Heike Zinsmeister
CLARIN-D PhD days – Corpora, Stuttgart, 26/03/2013
http://fr46.uni-saarland.de/lsteich/ClarindDS2013/
- Syntactic annotation in WebLicht (parts of speech, constituency, topological fields, grammatical functions)
- Search on syntactic annotated corpora (reference corpora and WebLicht output)
- Visualization and simple statistic analysis of search results

Tutorial: ELAN – Multimedia Annotator Alexander König
52. Studentische Tagung Sprachwissenschaften (http://www.stuts52.de/)
November 21–25, 2012, Berlin
With ELAN a user can add an unlimited number of annotations to audio and/or video streams. 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 provides several different views on the annotations, each view is connected and synchronized to the media playhead.
Up to 4 video files can be associated with an annotation document. Each video can be integrated in the main document window or displayed in its own resizable window.
http://tla.mpi.nl/tools/tla-tools/elan

Tutorial: Arbil – Metadata Editor Alexander König
52. Studentische Tagung Sprachwissenschaften (http://www.stuts52.de/)
November 21–25, 2012, Berlin
Arbil is an application for arranging research material and associated metadata into a format appropriate for archiving. It is designed so that it can be used off-line in remote locations. The data can be entered at any stage in part or as a whole. Arbil supports both IMDI, a metadata format developed within the scope of the DoBeS project at the MPI, and CMDI, the metadata format used within the pan-European CLARIN initiative. If used in CMDI mode, Arbil is built to integrate with the other parts of the CMDI framework, for example, the Component Registry and the ISOcat data category registry.
http://tla.mpi.nl/tools/tla-tools/arbil
LEXUS – Multimedia Lexicon Tool

Alexander König

November 21–25, 2012, Berlin

Lexus’s main target group are linguists who are involved in documenting endangered languages, but can be used as a lexicographical tool for any other language. Lexus does not enforce any pre-described schema structure to the user, it rather allows users to define the structure of their lexicon themselves. It deals with the interoperability problem by providing support for both the Lexical Markup Framework (LMF) and the ISO Data Category Registry (ISOcat).

Another distinct aspect of Lexus is the possibilities it offers for visualizing language. Its support for multimedia allows for local or archived images or movie clips to be attached to different parts of a lexical entry. Its support for customizable views of a lexical entry allows for different ways of styling lexicon for different purposes.

Finally, and most importantly, Lexus allows shared access to a given lexicon. This feature is fully adjustable by the creator of a given lexicon through read/write accessibility, and paves the way for better collaboration among researchers. Furthermore, it facilitates concurrent work on a given language by allowing multiple users access to the lexicon.

http://tla.mpi.nl/tools/tla-tools/lexus

Tutorial: ISOcat tutorial
21.05.2012, Istanbul
http://www.isocat.org/2012-LREC-ISOCat/

Tutorial: CMDI tutorial
05/06/2012, Oslo
http://clarin.b.uib.no/eventsevenementer/metadata-workshop-oslo-june-5-2012/
13/09/2012, Nijmegen
http://www.clarin.eu/node/3536

Tutorial: Metadata curation
26/10/2012, Sofia
http://www.clarin.eu/events/3539

Tutorial: Federated Content Search tutorial
24/04/2013, Copenhagen
http://www.clarin.eu/events/3579

Tutorial: Center Assessment tutorial
25/04/2013, Copenhagen
http://www.clarin.eu/events/3579

Tutorial: Einführung in die Tübinger LRT und Hilfestellung bei deren Anwendung für den Tübinger SFB 833
Wednesday, 19/1/2013

Tutorial: SpeechRecorder: a flexible recording software for scripted interviews
Christoph Draxler (BAS)
52. Studentische Tagung Sprachwissenschaften (http://www.stuts52.de/)
In linguistic fieldwork, e.g., in dialectology, sociolinguistics or phonetics, speech data is commonly collected in a standardized interview. SpeechRecorder is an application software for scripted audio recordings. Its main features are platform independence, multi-channel recording, the use of text, image or audio prompts to elicit speech, full support for Unicode, and different screens for speakers and recording supervisors. Recordings can be performed automatically or manually, and the recording items are selected sequentially or in random order. Recording phases allow fine control over item presentation and the beginning and end of a recording, and each recorded item is immediately saved to an audio file of its own, reducing the need for postprocessing.

SpeechRecorder is freely available from the Bavarian Archive for Speech Signals at LMU Munich as part of the CLARIN-D infrastructure initiative:

http://www.phonetik.uni-muenchen.de/forschung/Bas/software

Tutorial: Percy: online perception experiments for linguistics and phonetics
Christoph Draxler (BAS)
52. Studentische Tagung Sprachwissenschaften (http://www.stuts52.de/)
November 21–25, 2012, Berlin

Empirical research in linguistics and phonetics is often based on perception experiments. Performing such experiments online has a number of advantages over the traditional experiment in the lab: much larger audiences can be reached easily, experiments can be run parallel, participants may take part in the experiment at home, the experimenter bias is reduced and no special hardware or software is needed—experiments can be run on any modern web browser.

Percy is a software framework for online perception experiments. Unlike other web-based solutions, it is entirely written in HTML5 and JavaScript and thus runs on both traditional computers, handheld devices, e.g., smartphones or tablets, and Internet TVs. This provides access to entirely new target audiences. All experiment data is stored in a relational data-base system on the server, thus facilitating data exchange with spreadsheets or statistical software.

As part of the CLARIN-D infrastructure initiative, the Bavarian Archive for Speech Signals offers to host online experiments based on Percy free of charge to academic institutions. An overview of available experiments is available at:

http://webapp.phonetik.uni-muenchen.de/WebExperiment/

Tutorial: WebMAUS - how to automatically segment and label speech signals over the web
Thomas Kisler (BAS)
52. Studentische Tagung Sprachwissenschaften (http://www.stuts52.de/)
November 21–25, 2012, Berlin

Segmentation and labeling of speech signals is one of the basic tasks when confronted with research on spoken language. When done manually this task is time-consuming, error-prone and tedious.

The Munich Automatic Segmentation and Labeling System (MAUS) is an answer to this problem, as it allows the automation of this process. WebMAUS, an extension of MAUS, allows the user to access the underlying functionality over an easy-to-use web interface. For the automatic segmentation and labeling only a speech signal in wav-format and an orthographic transcript are necessary. The orthographic transcript is transformed to a phonetic transcription using BALLOON. These phoneme sequences are then passed to MAUS where they not only get forced aligned to the signal, but are aligned using a statistical model that represents the most likely production of an utterance.

The tutorial gives a basic overview of how MAUS and WebMAUS work and how the automatic
segmentation and labeling can be produced. The tutorial aims primarily at students and researchers with a phonetic background and/or anyone that is concerned with segmentation and labeling of audio data, everybody else interested in this topic is of course highly welcome.

**Tutorial: Introduction to EXMARaLDA**

The workshop introduced EXMARaLDA (“Extensible Markup Language for Discourse Annotation”), a system of concepts, data formats, and tools for the computer assisted transcription and annotation of spoken language, and for the construction and analysis of spoken language corpora. To find out more about EXMARaLDA please visit [http://www.exmaralda.org/index.html](http://www.exmaralda.org/index.html)

June 2012, Universität Augsburg
November 2012, Universidade do Minho, Braga
November 2012, Corpus Linguistics in the South 4, University of Portsmouth
November 2012, StuTS-Tagung, Berlin
December 2012, IDS, Mannheim
January 2013, Institut für Kulturologie und Anthropozentrische Linguistik an der Universität Warschau
February 2013, Ateliers méthodologiques, Université catholique de Louvain

**Tutorial: Query in Text Corpora**

Andreas Witt (Institut für Deutsche Sprache Mannheim)
“Culture & Technology”, European Summer School in Digital Humanities
[http://www.culingtec.uni-leipzig.de/ESU_C_T/node/38](http://www.culingtec.uni-leipzig.de/ESU_C_T/node/38)

**Course: “Digitale Editionsmethoden am Beispiel der Erbauungsliteratur des 17. Jahrhunderts”**

Susanne Haaf (DTA/CLARIN-D) / Matthias Schulz (DTA)
Freie Universität Berlin, Institut für Deutsche und Niederländische Philologie, Interdisziplinärer Masterstudiengang Editionswissenschaft
Wintersemester 2012/13

The goal of the course was to impart knowledge about methods for digitalizing editions using the example of funeral sermons from the 17th and 18th century. Within the framework of the DFG-Project AEDit the German Text Archive in cooperation with the Forschungsstelle für Personalschriften (Philippus-Universität Marburg) published 191 funeral sermons. Using selected examples of these widely published writings, the single steps necessary for digital editions of historical texts were taught: determining the edition criteria, creation of a digital primary text (transcription) and the annotation using XML/TEI as well as the commenting (cf. [http://www.fu-berlin.de/vv/lv/67235?m=189226&pc=74927&sm=11907](http://www.fu-berlin.de/vv/lv/67235?m=189226&pc=74927&sm=11907)). At the end of the course groups of 2–3 people created a digital edition of a funeral sermon according to the DTA transcription guidelines ([www.deutschestextarchiv.de/doku/richtlinien](http://www.deutschestextarchiv.de/doku/richtlinien)) and the DTA base format ([www.deutschestextarchiv.de/doku/basisformat](http://www.deutschestextarchiv.de/doku/basisformat)). Preliminary results (February 2013) could be made available via the DTA quality assurance DTAG ([www.deutschestextarchiv.de/doku/dtag](http://www.deutschestextarchiv.de/doku/dtag)).

**Course: Advanced Resources for Natural Language Processing.** Module A: Grammars and Tree-banks for Syntactic Processing

Heike Zinsmeister (co-taught with Stefanie Dipper, Ruhr Universität Bochum)
Nepal Summer School in Advanced Language Engineering, Dhulikhel, 27/08–07/09/2012
[http://nssnlp.ku.edu.np](http://nssnlp.ku.edu.np)

Creating and using resources for syntactic processing (including WebLicht):
• Grammars and their use in parsers
• Annotated corpora—constituency- and dependency-based treebanks
• Evaluation measures for inter-annotator agreement and system evaluation.

Besides, those centers who are involved in university teaching include CLARIN-D resources and tools in their courses on a regular basis.

4.3 Demos and Presentations

Demo: WebLicht
Kerstin Eckart, Heike Zinsmeister
Tag der Wissenschaft, Universität Stuttgart, 30/06/2012
General link: http://www.uni-stuttgart.de/tag/2012/
We presented WebLicht at the Stuttgart University’s Science Day to other researchers and the interested public. We exemplified the use of WebLicht by sample cases (e.g., Named Entity recognition), and showed that WebLicht allows the user to employ tools that are located at different centers.

Presentation: CLARIN D—creating an infrastructure for language-related research
Heike Zinsmeister
Institutsversammlung, Universität Stuttgart, 16/05/2012

Presentation: Projekte im Bereich eHumanities: CLARIN-D, e-Identity, ePoetics
Jonas Kuhn
Themenabend zu eHumanities, Forschungsverbunds Sprachwissenschaft und Kognition, Universität Stuttgart, 24/05/2012
http://www.fsk.uni-stuttgart.de/Archiv/FSK-Workshops_neu/Workshop_15.html

Presentation: Technical aspects of the CLARIN-D infrastructure
Jens Stegmann
IWiST-Institutsversammlung, Universität Hildesheim, 06/11/2012
http://www.uni-hildesheim.de/index.php?id=iwist

• Repository: Fedora Commons, PIDs, OAI-PMH
• CMDI metadata

Presentation: Interaktive Textanalyse und das CLARIN-D-Infrastrukturprojekt
Andre Blessing, Kerstin Eckart, Jonas Kuhn, Jens Stegmann, Heike Zinsmeister
Ringvorlesung Digital Humanities: Projekte und Perspektiven, Universität Stuttgart, 18/12/2012
http://julienas.philosophie.uni-stuttgart.de/groups/ehumanities/

• CLARIN-D project goals
• Metadata and retrieval of resources

13
• Preprocessing and machine annotation in WebLicht (incl. demo)
• Manual annotation (EXMARaLDA, WebAnno),
• Speech segmentation and annotation (WebMaus)
• Interactive text analysis (incl. demo)

Presentation: “eTRACES - Winged words, quotations and our cultural heritage”
Marco Büchler (Universität Leipzig)
“Culture & Technology”, European Summer School in Digital Humanities
http://www.culingtec.uni-leipzig.de/ESU_C_T/node/76

Presentation: “Research Infrastructures at the Institute for the German Language”
Andreas Witt (Institut für Deutsche Sprache Mannheim)
“Culture & Technology”, European Summer School in Digital Humanities
http://www.culingtec.uni-leipzig.de/ESU_C_T/node/76

Presentation: “Legal issues and best practices for multimodal resources.”
Erik Ketzan, Pawel Kamocki (Institut für Deutsche Sprache Mannheim)
Universität Bielefeld, September 21, 2012

4.4 Other

coordination: Axel Herold (BBAW)
The CLARIN-D User Guide is meant as a best practice for the adaptation and integration of existing language resources into the CLARIN-D infrastructure. The first part deals with basic thematic issues and modeling processes, which are applicable to all kinds of linguistic resources and tools (data categories, metadata, annotation, legal issues, quality management). The second part deals with resource specific characteristics of the infrastructure (corpora, lexical resources, linguistic tools) with a specific focus on the description of technical specifics. The creation of the User Guide was a collaborative work of members from different CLARIN-D partner institutes: Kathrin Beck (Univ. Tübingen), Kerstin Eckart (Univ. Stuttgart), Alexander Geyken (BBAW Berlin), Lothar Lemnitzer (BBAW Berlin), Susanne Haaf (BBAW Berlin), Axel Herold (BBAW Berlin), Erik Ketzan (IDS Mannheim), Marc Kupietz (IDS Mannheim), Harald Lüngen (IDS Mannheim), Scott Martens (Univ. Tübingen), Ingmar Schuster (Univ. Leipzig), Dieter van Uytvanck (MPI Nijmegen), Andreas Witt (IDS Mannheim), Peter Wittenburg (MPI Nijmegen), Thomas Zastrow (Univ. Tübingen), Heike Zinsmeister (Univ. Stuttgart). The CLARIN-D User Guide was published on December 19, 2012 on the CLARIN-D website: http://de.clarin.eu/en/language-resources/userguide.html A HTML version as well as a PDF document are available, an EPUB format is in preparation.

Panel Discussion: “Virtual Research Infrastructures in the (Digital) Humanities”
“Culture & Technology”, European Summer School in Digital Humanities
http://www.culingtec.uni-leipzig.de/ESU_C_T/node/154
The first panel discussion open to the public will be devoted to the question: “Which directions have to be taken if Virtual Research Infrastructures are to serve the (Digital) Humanities”? The reason why this question is being asked is that in discussions around Virtual Research Infrastructures, when it comes to the Humanities, “for the Humanities” in the sense of “Computer Science develops virtual research infrastructures for the Humanities” would seem to be more central than “Humanities” in the sense of “Virtual Humanities Research Infrastructures” and thus the collaboration on an equal level of the Humanities and Computer Science. The panel discussion aims at the initiation of a reflection on these two different perspectives and their implications.
5 Cooperation

5.1 Cooperation with AP7 Support & Helpdesk

In order to be able to assist and help potential users in an effective way, a close cooperation between AP7 (Support & Helpdesk) and AP8 (Training & Education) is essential.

The helpdesk platform, TeLeMaCo, training and hosting measures need to be interlinked bidirectionally to provide users with as much information and/or assistance needed. We can imagine two scenarios:

- inquiry to the helpdesk may result in
  - links to relevant material in TeLeMaCo
  - information about tutorials/workshops
  - information about the possibility of hostings

- questions arising from a query for teaching and learning material (TeLeMaCo)
  - link to helpdesk platform
  - FAQs

Besides, the questions addressed to the helpdesk can be used to determine needs of users for further teaching and training material.

AP7 and AP8 conducted a joint workshop in order to discuss potential, questions and problems in this regard. The program of the workshop can be found in the appendix.

5.2 Cooperation with DARIAH AP2

The CLARIN-D team in Saarbrücken took part in two meetings of the WP2 of the DARIAH project. The aim of the DARIAH-DE project (http://de.dariah.eu/) is to support and extend the use of digital methods in the humanities. DARIAH-DE WP2 is concerned with research and education and is thus the counterpart work package to AP8 in CLARIN-D.

The first DARIAH-DE WP2 meeting took place in Würzburg at the Julius-Maximilians-Universität Würzburg on the 25th of May 2012. During these meetings the CLARIN-D representative, Dr. Michaela Vela, presented the CLARIN-D project with a special focus on WP8. The agenda in Würzburg included the following topics:

- ontology and typology of the research processes as well as demonstrators
- DH definition, DH brochure
- training sessions and workshops (state of affairs, ideas for further events)
• perspectives for the cooperation with CLARIN

• representation of the conceptual work from WP2 during the meeting of the scientific advisory board

• work schedules for the second year (progress, coordination need), ideas for the subsequent application

• VCC2: Virtual Competency Centre Research and Education

The second DARIAH-DE meeting took place at Technische Universität Darmstadt on the 10th of December 2012. At this meeting the CLARIN-D representatives, Prof. Elke Teich and Dr. Michaela Vela, presented the Linguistic Teaching Resources Hub (TeLeMaCo), which was implemented in Saarbrücken as part of WP8 in CLARIN-D, and took part in the discussions concerned with the following topics:

• DARIAH-DE further educational courses,

• DARIAH-DE portal and interactive graphic overview for the portal,

• state of the considerations for the publication service (report CS and StS): [Series see DARIAH Working papers]; technical putting into action, preparations for the contents,

• Collection registry and Geobrowser,

• VCC2 (Virtual Competency Centre) and report from the Nedimah conference in London,

• cooperation with CLARIN-D—presentation of the CLARIN-D repository for teaching and learning material TeLeMaCo,

• subsequent application DARIAH-DE: restructuring and orientation,

• presentation of the questionnaire outline “feedback for demonstrators”,

• curricular developments.
6 Appendix
Abstract

Naturally-occurring text in many languages are annotated for linguistic structure. A Treebank is a text corpus in which each sentence has been annotated with syntactic structure. Treebanks are often created on top of a corpus that has already been annotated with part-of-speech tags. The annotation can vary from constituent to dependency or tecto-grammatical structures. Treebanks have become valuable resources as repositories for linguistic research.

In this report we describe our experimental undertaking on building parallel Treebanks for German-Georgian, German-Russian and German-Ukrainian language pairs. The languages (except German) involved in the project from the computational viewpoint are considered “lesser-resourced” languages.

The parallel Treebanks can be used in translation studies, in corpus linguistics for studying syntactic phenomena, in computational linguistics as evaluation corpora for different NLT systems or for training and testing parsers and as a database for Translation Memory systems.

1. Introduction

Parallel corpora are language resources that contain texts and their translations, where the texts, paragraphs, sentences, and words are linked to each other. In the past decades they became useful not only for NLP applications, such as machine translation and multi-lingual lexicography, but are considered also very useful for empirical language research in contrastive and translation studies.

Naturally-occurring text in many languages are annotated for linguistic structure. A Treebank is a text corpus in which each sentence has been annotated with syntactic structure. Treebanks are often created on top of a corpus that has already been annotated with part-of-speech tags. The annotation can vary from constituent to dependency or tecto-grammatical structures. In turn, Treebanks are sometimes enhanced with semantic or other linguistic information and are skeletal parses of sentences showing rough syntactic and semantic information.

Treebanks have become valuable resources as repositories for linguistic research. They can be used in translation studies, in corpus linguistics for studying syntactic phenomena, in computational linguistics as evaluation corpora for different NLT systems or for training and testing parsers.

In this report a work on building parallel Treebanks for the language pairs German-Georgian, German-Russian and German-Ukrainian is outlined. From the computational viewpoint three of the mentioned languages, except the German Language, are considered to be “the lesser-
resourced languages”. Besides, typologically German and Georgian is much more dissimilar language pair, than the rest two pairs are.

The objective of the mentioned visit was not development of the full-scale parallel treebanks for the three languages pairs which would be unrealistic given the short notice of the research stay at the University of Saarland. Rather, the aim was starting with simple sentences in all four languages

- to tag and lemmatize manually terminal nodes
- produce syntactic parses for monolingual parallel German, Georgian, Russian and Ukrainian resources
- compare nonterminal nodes for determining stable and possible equivalents between phrases across the syntactic structures of the languages involved
- establish compatible tag-sets for Georgian, Russian and Ukrainian and, if necessary, to introduce the new syntactic phrase categories.

On the ground of the developed monolingual resources the further objective of the experiment envisioned

- production of the parallel trees for the bilingual resources
- alignment of the German-Georgian, German-Russian and German-Ukrainian parallel trees
- making general conclusions concerning feasibility of development treebanks for the mentioned language pairs.

2. Resources for experiment

For the low-density languages, including Georgian, Russian and Ukrainian, parallel corpora are very rare. The parallel texts used for the outlined experiment comprises German sentences and their translations into Georgian and Russian languages compiled for the GREG NLP lexicon project (Kapanadze et al., 2002, Kapanadze, 2010). The GREG itself contains valency data with the manually aligned Georgian, Russian, English and German verbs (ca. 1250) augmented with the examples of sentences considered as translation equivalents. Each subcorpus used for the study has a size of roughly 2600 sentence pairs that correspond to different syntactic subcategorization frames considered as German-Georgian translation equivalents. For the Russian and Ukrainian languages translation equivalents were provided by Dr. Alla Mishchenko, a DAAD postdoctoral fellow at the University of Saarland. She also took an active part in development of the monolingual Russian and Ukrainian resources and alignment procedures of the bilingual German-Russian and German-Ukrainian treebanks.

3. Building Monolingual Treebanks

3.1. Morphological analysis.

Initially emphasis has been made on development of a parallel treebank for a typologically dissimilar language pair German and Georgian, since the later is an agglutinative language using both suffixing and prefixing. For the Georgian text analyses has been applied a finite-state morphological transducer using the XEROX FST tools (Kapanadze 2010a,b), (Kapanadze 2009). The Georgian FST transducer utilizes a number of the formalisms supported by the XEROX toolkit (Beesley and Karttunen, 2003). The lexicon specification language lexc was used for modeling the lexicon and for constraining the morphotactics. It consists of 7 modules
for noun, adjective, pronoun, numeral, adverb, verb and the minor categories analysis. Currently there are two versions of the Georgian FST transducer available in the MS Windows platform and in the LINUX UBUNTU version.

For the rest of languages, German, Russian and Ukrainian, involved in the experiment, morphological features, including POS tags, were assigned manually drawing on the TIGER guidelines for the German language with the necessary changes relevant to the Russian and Ukrainian grammar formal description.

3.2 Syntactic parsing

The syntactic annotation employs parts-of-speech tags, morphological properties, and dependency functions. Every sentence is assumed to have a unique head and all other tokens, except punctuation marks, are direct or indirect dependents of the head. Monolingual files are XML-formatted.

Using the morphologically annotated bilingual corpus for each pair (German-Georgian, German-Russian and German-Ukrainian) the syntactical annotation were done manually. For this purpose we utilised the Synpathy, a tool for syntactical annotation developed at Max Plank Institute for Psycholinguistics, Nijmegen, the Netherlands (www.mpi.nl/corpus/manuals/manual-synpathy.pdf), a CLARIN-D project collaborator.

The German treebank annotation follows the TIGER annotation scheme (Skut et al., 1997, Brants et al., 2002). The other three monolingual treebank were annotated according an adapted version of the German TIGER guidelines. The output of the syntactic annotation is in the TIGER-XML format. From the TIGER-XML format, the syntactic annotation may be visualized with tools like TIGER Search, representing dependency graphs for sentences. In Figure 1 and Figure 2 are examples of dependency trees for German, Georgian and Russian Sentences.

![Figure 1: A screenshot of an annotated sentence in German language.](image-url)
Figure 2: A screenshot of the corresponding annotated Georgian sentence.

Figure 3: A screenshot of the corresponding annotated Russian sentence.
The monolingual treebanks converted into TIGER-XML, are a powerful database-oriented representation for graph structures. In a TIGER-XML graph each leaf (= token) and each node (= linguistic constituent) has a unique identifier (Samuelsson and Volk, 2007). We use these unique identifiers for the phrase and word alignment across trees in corresponding translation units.

An XML representation is also used for storing this alignment. In the Figure 5 there is a representation of the Georgian sentence from the Figure 2 in the TIGER-XML format.

```xml
<body>
  <s id="s12">
    <graph root="s12_502" discontinuous="true">
      <terminals>
        <t id="s12_1" word="ის" pos="DPRN" morph="Nom.3.Sg" />
        <t id="s12_2" word="აგიტაციას" pos="NN" morph="Dat.Sg" />
        <t id="s12_3" word="ეწეოდა" pos="VVFIN" morph="Sb3.Sg.Ob3.Pret" />
        <t id="s12_4" word="ამ" pos="DPRN" morph="Gen.Sg" />
        <t id="s12_5" word="მთავრობის" pos="NN" morph="Gen.Sg" />
        <t id="s12_6" word="წინააღმდეგ" pos="PPS" morph="Gen" />
        <t id="s12_7" word="." pos=".$" morph="--" />
      </terminals>
    </graph>
  </s>
</body>
```
4. Building Parallel Treebanks.

Alignment of a Monolingual German, Georgian, Russian and Ukrainian Treebanks into a Parallel Treebank

This procedure is done with the help of the Stockholm TreeAligner, a tool for work with parallel treebanks which inserts alignments between pairs of syntax trees (Samuelsson and Volk, 2005, Samuelsson and Volk, 2006). The Stockholm TreeAligner handles alignment of tree structures, in addition to word alignment, which – according to its developers - is unique (Samuelsson and Volk, 2006).

Phrase alignment can be regarded as an additional layer of information on top of the syntax structure. It shows which part of a sentence in the German language is equivalent to which part of a corresponding sentence in the other language. This is done with the help of a graphical user interface of the Stockholm TreeAligner. We drew alignment lines manually between pairs of sentences, phrases and words over parallel syntax trees. Figure 6 shows a screenshot with two aligned trees from Figure 1 and Figure 2. We intended to align as many phrases as possible. The goal is to show translation equivalence. Phrases shall be aligned only if the tokens, that they span, represent the same meaning and if they could serve as translation units outside the current sentence context. The grammatical forms of the phrases need not fit in other contexts, but the meaning has to fit.

The Stockholm TreeAligner guidelines allow phrase alignments within m : n sentence alignments and 1 : n phrase alignments. Even though m : n phrase alignments are technically possible, we have only used 1 : n phrase alignments, for simplicity and clarity reasons. One example of 1: n alignment on the word level is the Georgian multi-word expression for
“აგიტაციის გაწევა” represented under a VP node in the Figure 2, which is one word ("agitierte") in the corresponding German sentence from the Figure 1. The 1 : n alignment option is not used if a node from one tree is realized twice in the corresponding tree, e.g. a repeated subject in coordinated sentences.

The Stockholm TreeAligner differentiates between two types of alignment, displayed by different colours. Nodes and words representing exactly the same meaning are aligned as exact translation correspondences using the green colour for lines as it is shown on the Figure 6. In this regard a German word ("agitierte") alignment to the Georgian Verb Phrase “აგიტაციის გაწევა” as an exact one, might be considered problematic.

Figure 6: A screenshot with aligned trees from Figure: 1 and Figure: 2.
Nevertheless, in such a case a prerequisite for this solution is that they could serve as translation units outside the current sentence context. If nodes and words represent just approximately the same meaning, they are aligned as fuzzy translation correspondences by means of lines in the red colour as it is shown in the Figure 7 above.
In an appendix an example of an annotated compound Georgian and German sentences with exact and fuzzy alignment on simple clause and phrase level could be viewed.

5. Conclusions

At the initial phase of presented experiment we made an overview of experience in building parallel threebanks for languages with different structures (Megyesi and Dahlqvist, 2007), (Megyesi et al., 2006), (Grimes et al., 2011), (Rios et al., 2009).

As it is reported in a Quechua-Spanish parallel treebank project, due to strong agglutinative structure of the Quechua language, it was decided to annotate the Quechua treebank on morphemes rather than words. This allowed the authors to link morpho-syntactic information precisely to its source. Besides, building phrase structure trees over Quechua sentences does not capture the characteristics of the language. Therefore, they have chosen Role and Reference Grammar. By using nodes, edges and secondary edges in the Stockholm annotation tool they were able to represent the most important aspects of Role and Reference syntax for Quechua sentences (Rios et al. 2009).

Although the Georgian language is also an agglutinative language with suffixing and prefixing, there is no need to annotate the Georgian Treebank on morphemes. However, for syntactic annotation in the Georgian language a precise description of a specific structure/mechanism of its clause is necessary. “The Georgian clause is a word collocation which draws on coordination and government of the linked verb and noun sequence” [Chikobava, 1928]. The types of
syntactic relations in the Georgian clause differ significantly from those observed in the Indo-European or in other languages. In the English Language there are just a small number of verbs that govern the nouns linked to them as indirect actants and demand those nouns to stand in an indirect case form (e.g. John believes him to be innocent). Besides, the actants involved do not induce changes in the verb form. In contrary, in the polyvalent Georgian verb the actants are marked with specific affixes in a verb. The most significant difference from the structure of the Indo-European syntactic relations model is that in the Georgian clause there is a mutual government and agreement relations or a bilateral coordination phenomenon between verb-predicate and noun-actants which number may reach up to three in a single clause. It anticipates control of the noun case forms by verbs, whereas the verbs, in their turn, are governed by nouns with respect to a grammatical person. Therefore, according to [Chikobava, 1928] in a syntactic description of Georgian the concepts of a Major and a Minor Coordinate, instead of Subject and Object, are preferable. Moreover, in the verb forms of a certain semantic type an indirect object has preference as a Major Coordinate over a Subject (a Minor coordinate) in the respect of its marking in a verb form. Nevertheless, unlike the Quechua language, Georgian syntax can be sufficiently well represent by means of dependency relations and there is no need to utilize a different approach to capture the Georgian language structural peculiarities.

The Russian and Ukrainian languages typologically are more closely related languages to German than Georgian is. Consequently, tag-sets for these two languages underwent minor changes and some additional POS and CAT features has been introduced. The changes for the Georgian language tag-sets and CAT values are more significant, but in general they conform to the TIGER guidelines which served as a background in compiling the features and their values for all three new languages involved in the project.

Besides, the Georgian, Russian and German languages also fairly good conform to the TIGER-xml format and syntactic trees perfectly reflect skeletal parses for each those languages.

The TIGER-XML format (.tig extension) is the treebank exchange format allowing free data exchange and the use of tools developed by the international TIGER project community. In the TIGER format, edge labels contain the original syntactic function tags, and the (non-terminal) cat category contains phrase and clause forms. A TIGER-XML file consists of a header and a body. The corpus header can contain meta-information about the corpus (such as corpus name, date, author, etc) and a declaration of the tags that are used in the morphology Part-of-Speech, non-terminal nodes and edges. In the second part of a TIGER-XML file the corpus body contains words, Part-of-Speech tags, morphological tags and lemmata which are listed as attributes of the element "terminal". Non-terminals are represented in an additional element called "nonterminal" referring to the corresponding terminal ID. This part of the XML file contains the encoding for secondary edges as well.

A big advantage of using the xml format is exchangability and usability with a large range of other applications. For example the TIGERsearch corpus query tool, and in the multimedia annotation tools as ELAN and ANNEX.

**Monolingual resources** with .tig extension for browsing by the Sympathy tool are in the appended folders as followes:

- GER – for sentences in the German language;
- GEO – for sentences in the Georgian language;
- RUS – for the sentences the Russian language;
- UKR– for the sentences the Ukrainian language.
Bilingual aligned sentences in .xml format for browsing by means of the Stockholm TreeAligner are in the following folders:

AGEGO – German-Georgian;
AGERU – German-Russian;
AGEUK – German-Ukrainian.

**ADDENDUM:**

1. Examples of aligned German-Georgia, German-Russian and German-Ukrainian.


**References**


Day 1   Mo 25.3.2013
13:30 - 17:00   Corpus query with CQP (Hannah Kermes, UdS):
               token-based queries,
               regular expressions,
               simple "statistical" analysis (grouping)

ca. 15:00 - 15:30  Pause/break

Day 2   Tue 26.3.2013
09:00 - 12:30   Syntactic annotation and corpus search with CLARIN-D resources (Heike Zinsmeister, IMS)
               Syntactic annotation in WebLicht (parts of speech, constituency, topological fields, grammatical functions)
               Search on syntactic annotated corpora (reference corpora and WebLicht output)
               Visualization and simple statistic analysis of search results

ca. 10:30 - 11:00  Pause/break
12:30 - 13:30  Mittagspause/lunch break
13:30 - 17:00   Statistical analysis with R (Marilisa Amoia, UdS)

ca. 15:00 - 15:30  Pause/break
Support und Helpdesk-Workshop (27.11)

10:00-10:15  **Begrüßung**  
(Timm Lehmberg)

10:15-10:45  **Beratung und Support für E-Humanities - Ein Erfahrungsbericht**  
(Frank Binder)

10:45-11:15  **CLARIN-D Schulung und Ausbildung**  
(Hannah Kermes)

11:15-11:45  **Die CLARIN-D [Informationsplattform](#) und der Legal Helpdesk**  
(Erik Ketzan, Pawel Kamocki)

11:45-12:00  **Das CLARIN-D Benutzerhandbuch**  
(Axel Herold)

12:00-13:00  **Mittagspause**

13:00-13:30  **Der CLARIN-D Helpdesk**  
(Timm Lehmberg)

ab 13:30  **Diskussion**

- Vernetzung der AP
- Integration der AP in das Ticketing System
- Vernetzung von Wissensressourcen
- M24 Demonstrator
- ...

**Teilnehmende:** Kathrin Beck (Tübingen) Frank Binder (Gießen) Volker Boehlke (Leipzig) Pawel Kamocki (Mannheim) Yael Dilger (Hamburg) Fideniz Ercan (Hamburg) Kai Wörner (Hamburg) Erhard Hinrichs (Tübingen) Axel Herold (Berlin) Daniel Jettka (Hamburg) Hannah Kermes (Saarbrücken) Erik Ketzan (Mannheim) Jörg Knappen (Saarbrücken) Timm Lehmberg (Hamburg) José Martínez (Saarbrücken) Beatriz Sanchez (Garching) Thomas Kisler (München) Thomas Schmidt (Mannheim) Jens Stegmann (Stuttgart) Daniel Stein (Hamburg) Elke Teich (Saarbrücken) Thomas Zastrow (Tübingen) Heike Zinsmeister (Stuttgart)
Bibliography


