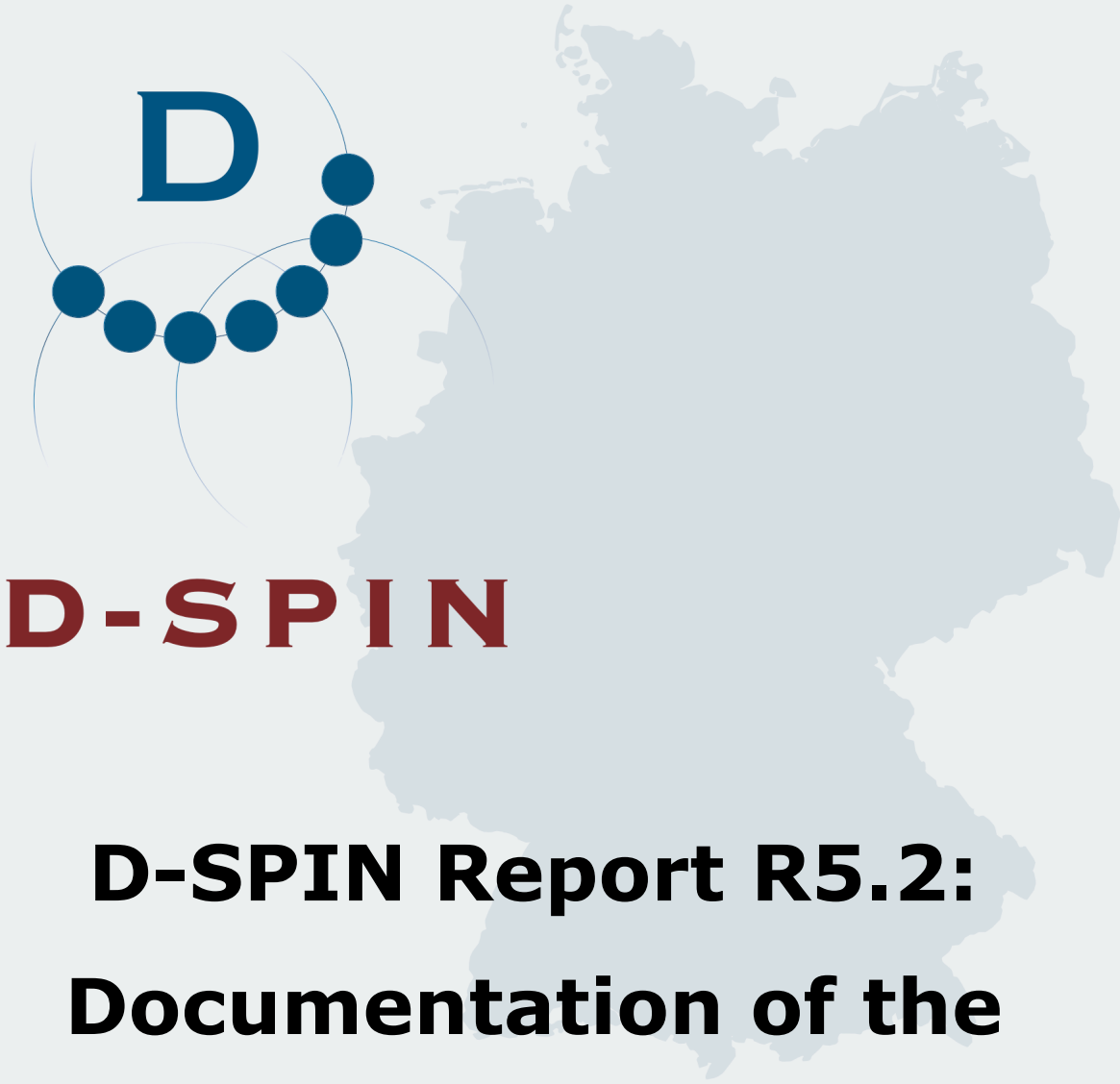

D-SPIN

D-SPIN Report R5.2: Documentation of the D-SPIN preparation activities

April 2010

D-SPIN, BMBF-FKZ: 01UG0801A

Deliverable:R5.2: Documentation of the D-SPIN preparation activities

Responsible: Erhard Hinrichs

Editors: Erhard Hinrichs, Kathrin Beck

Contributors: Volker Boehlke, Thierry Declerck, Ulrich Heid, Jost Gippert, Marc Kupietz, Lothar Lemnitzer, Peter Wittenburg,

1. Introduction

The D-SPIN report R5.1 “Guidelines for the Standard-Conformant Adaptation and Creation of Language Resources” presents an overview of existing standards for LRT in CLARIN and ISO. These guidelines provide help by the choice of the most suitable standards.

For the integration of own language resources and tools, it is important to apply to uniform standards in order to be compatible with the LRT federation.

D-SPIN Report R5.2 “Documentation of the D-SPIN preparation activities” is a descriptive document of the LRT data preparation activities of the D-SPIN partners.

It provides an overview on the linguistic resources and that have been developed in previous and current projects. And on activities that have already taken place or that are planned, e.g. to adapt to CLARIN standards and/or to integrate resources and tools into the D-SPIN WebLicht platform.

2. Standards

D-SPIN is collaborating in the European standardization process of CLARIN in several domains. In addition to the international dialogue of the agreements on standards, own language resources and tools of D-SPIN partners are being adapted to these standards.

2.1 Metadata

Metadata are an important part of all LRT. They are information describing properties of linguistic resources, e.g. the size of a corpus, the recording date of a speech file, the purpose for which annotations were created. They are described in metadata schemata like Dublin Core, OLAC, IMDI, TEI, etc.

2.2 ISOcat

ISOcat (http://www.isocat.org/) is a Data category registry hosted and enhanced at the MPI Nijmegen. ISOcat is the software and database that implements the ISO 12620 standard and data model which is currently being filled with many categories from the EAGLES project, various metadata initiatives a.o. The ISO DCR is a cornerstone of the standardization and interoperability framework.

2.3 ISO formats

An overview of the here-mentioned ISO standards is provided in the D-SPIN Report 5.1 “Guidelines for the standard-conformant adaptation and creation of language resources”.

The D-SPIN WP 5 cooperates with Laurent Romary of the ISO committee TC 37/SC 4 on the development of the ISO standards LAF, MAF, and SynAF. Detailed specifications and definitions are being worked out, but there are no final versions yet. After that, they will be implemented on some tools of the WebLicht tool chain for evaluating their performance.

2.3.1 Linguistic Annotation Framework (LAF)

The Linguistic Annotation Framework (LAF) is a framework for linguistic annotation of language resources that can serve as a reference or pivot for different annotation schemes. With generalized annotation formats, sharing, merging, and comparison of language resources will be facilitated. LAF provides a generic framework for representing annotated resources as graphs and nodes and links associated to feature structures (conformant to ISO 24610).

2.3.2 Morpho-syntactic Annotation Framework (MAF)

The Morpho-syntactic Annotation Framework (MAF) offers a model for representing morpho-syntactic annotation a two-tier principle of token and word form. It offers a tag-set definition framework but doesn’t specify a unique tagset.

2.3.3 Syntactic Annotation Framework (SynAF)

The Syntactic Annotation Framework (SynAF) supplies a reference format for the representation of syntactic annotations. The model is general enough to represent both constituent and dependency based syntactic annotation.

2.4 TCF format

The Text Corpus Format (TCF) was developed in D-SPIN as internal data format of the WebLicht web services. It is a simple standoff format for linguistic annotated text corpora. The format is used in the WebLicht tool chain for sending partly annotated text corpora in an efficient way from the chaining mechanism to the individual web services.

The TCF-Format is defined via NG Relax and XML schema. Its architecture is open, so that with rising needs adjustment is possible. TCF is a multi-layer standoff format. In contrast to some other formats (for example PAULA or MAF), TCF stores all linguistic layers in one file. That means that during the chaining process, the file grows.

A plain text in the TCF format looks the following way:

```
<?xml version="1.0" encoding="UTF-8"?>
<D-Spin xmlns="http://www.dspin.de/data" version="0.3">
    <tns:MetaData xmlns:tns="http://www.dspin.de/data/metadata">
        <tns:source></tns:source>
    </tns:MetaData>
    <tns:TextCorpus xmlns:tns="http://www.dspin.de/data/textcorpus" lang="de">
        <tns:text>Karin fliegt nach New York. Sie will dort Urlaub machen.</tns:text>
    </tns:TextCorpus>
</D-Spin>
```

3. Available resources in D-SPIN

The D-SPIN partners have a large variety of language resources and tools. Most of them were developed on-site during research projects.

3.1 Max Planck Institute (MPI) Nijmegen

Language resources

The MPI houses a large archive for language resources of all data types from a variety of sub-disciplines such as: endangered languages, multimodal studies, multilingualism studies, child and adult language acquisition studies, sign language studies, brain imaging studies, etc. The archive currently covers 50 Terabyte of data (yearly increase about 10 TB), mostly of course used by the digital versions of video and sound files. The archive is open for other researchers, an opportunity that is heavily used in Germany and the Netherlands in particular.

In many cases we cannot speak about corpora anymore, the archive is a living body where researchers add new data or new versions of existing data according to their research needs. It covers currently about 60,000 files with textual material in different formats as the following table indicates. Mostly these files are annotations on media files, however there are also lexica and descriptive files.

In average one can say that there are 3 media files (sound, video archive format and video presentation format) per annotation file. However, the amount of media and time series files (eye tracking, eeg, fMRI, gesture, etc) files that are not analyzed and annotated is increasing.

	EAF	CHAT	SBX	Text	XML	HTML	total
MPI	4645	4836	551	5111	16	410	15569
DOBES	3509	136	324	284	59	60	4372
DBD	10	675		85	1	1	772
SignLang	2670						2670
ESF		1643		1145			2788
EL	4162	136	324	300	59	121	5102
IFA				15632			15632
CGN	12767						12767
total	27763	7426	1199	22557	135	592	59672

MPI: all data from MPI researchers; DOBES: all data from the DOBES endangered languages project; DBD: Dutch bilingualism project; SignLang: data from various sign language researchers in Europe; ESF: Adult language acquisition data from 5 EU countries; EL: material about endangered languages from various depositors worldwide; IFA: phonetic corpus; CGN: Dutch Spoken corpus;

All these files are accessible via the online metadata catalogue, which is based on IMDI metadata. It can be found under: http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI1%23

Since all metadata is harvestable via OAI-PMH all resources are also available now via the CLARIN Virtual Language World: http://www.clarin.eu/vlw/observatory.php

Language Archiving Technology

The MPI has developed the Language Archiving Technology software suite with different types of components:

- Components that work on local computers
- Components that work as web applications
- Components that can be invoked as services
- Components that can be used locally and as web applications

The following table gives an impression about the types of components.

	Function	web/ local	API available	CLARIN/ D-SPIN	State	launch
ELAN	Multimedia/multimodal annotation & analysis tool	local	no		mature	01
ANNEX	Multimedia/multimodal annotation & analysis tool	web	yes		mature	06
TROVA	Content search & analysis tool	web& local	planned		mature	06
LEXUS	Lexicon creation, analysis & visualization tool	web& local	yes		UI improvemen ts	06
VICOS	graphical conceptual space creation and browsing tool	web& local	yes		mature, new functions	07
IMDI Editor	editing of IMDI metadata descriptions	local	no		mature	01
IMDI Browser	browsing and viewing of IMDI hierarchies	web & local	no		mature	01
IMDI Infra	OAI PMH harvesting and offering, etc	web	yes		mature	02
Virtual Language Observatory	metadata portal technology with support of GIS, facetted browsing, mapping	web	yes	yes	in development	09
CMDI Infra	new CLARIN metadata infrastructure tools	web	in dev	yes	in development	in 10
ARBIL	new metadata editor and organizer, will be CMDI editor	web & local		yes	CMDI functionality in developm	09
LAMUS	repository system	web	yes		mature	05
AMS	access management system	web	yes		mature	05
COSIX	data synchronization software	web	no		testing phase	09
REPLIX	data replication based on iRODS	web	no	yes	in development	to come
ISOcat	data category editing and visualization on behalf of ISO	web	yes	yes	in use, additional functions	09

The web-based tools will all be equipped with APIs to allow other services to invoke them if this has not yet been done already. All tools are open source and freely available. ELAN is currently one of

the most widely used multimedia/multimodal annotation tools worldwide. Where possible all tools support international standards such as UNICODE, XML, OAI PMH, ISO 12620 etc.

3.2 University of Tübingen

The Department of Computational and Applied Linguistics at the University of Tübingen hosts a variety of language resources, in particular corpora that are annotated manually. GermaNet and TüBa-D/Z are long-term projects that are continuously enlarged.

Lexical resources

Wordnets

- **GermaNet** – GermaNet is a lexical-semantic German word net, which provides for some 53,000 general language concepts (so-called synsets) with more than 76,000 word meanings and the most important semantic relations which hold among represented concepts and word senses. GermaNet covers the German base vocabulary. The core subset of GermaNet (1 300 Base Concepts) is specified in terms of the EuroWordNet Top Ontology, a system consisting of 63 semantic features which are similar to semantic universal categories like semantic primitives.

Corpora

- **Tübingen Treebank of Spoken German (TüBa-D/S) –** The TüBa-D/S is a syntactically annotated corpus based on spontaneous dialogues which were manually transcribed. It was annotated as part of the Verbmobil project and comprises approximately 38,000 sentences (ca. 360,000 words). The syntactic annotation was performed manually.
- **Tübingen Treebank of Written German (TüBa-D/Z)** – The TüBa-D/Z is a syntactically annotated newspaper corpus based on data of they daily newspaper "die tageszeitung". It contains approximately 45,200 sentences with almost 800,000 tokens. The corpus is manually annotated and contains the following annotation layers: Inflectional morphology, syntactic constituency, grammatical functions, (complex) named entities, anaphora and coreference relations.
- **Tübingen Treebank of Spoken English (TüBa-E/S)** – The TüBa-E/S is a syntactically annotated corpus based on spontaneous dialogues, which were manually transcribed. It was annotated as part of the Verbmobil project and comprises approximately 30,000 sentences (ca. 310,000 words). The syntactic annotation was performed manually.
- **Tübingen Treebank of Spoken English (TüBa-J/S)** – The TüBa-J/S is a syntactically annotated corpus based on spontaneous dialogues, which were manually transcribed. It was annotated as part of the Verbmobil project and comprises approximately 18,000 sentences (ca. 160,000 words). The syntactic annotation was performed manually.
- **The Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)** – TüPP-D/Z is a collection of articles from the daily newspaper, "die tageszeitung", which have been automatically annotated with clause structure, topological fields, and chunks, in addition to more low-level annotation including parts of speech and morphological ambiguity classes.

The TüPP-D/Z data of the current release is taken from the 1999 HTML distribution (scientific edition) of the "tageszeitung", which includes newspaper articles from September 2, 1986 up to May 7, 1999 and which amounts to more than 200 million word tokens of text.

Tools

- **CASS** – CASS is a finite-state parser which uses a cascade of finite-state automata. It is optimized for speed and memory. CASS was developed by Steven Abney while he was in Tübingen. The latest version of this parser can be downloaded from his website.
- **BibTex-HTML Converter** – Our bibTeX-XML-HTML package transforms a (potentially thematically structured) BibTeX bibliography into a set of HTML files to facilitate their publication on the web.

  The BibTeX-XML-HTML converter tools will first transform your BibTeX file into an XML file. The advantage of the XML format is that it is standardized, platform independent and universal. Furthermore, it is easily edited and managed with XML tools and translated into HTML (or text files) with the XSLT style sheet language. Our tools will generate multiple HTML files out of the XML representation of the bibliography for publication on the web.
- **FSA Utilities** – The FSA Utilities are a collection of tools that construct finite automata from regular expressions, manipulate finite automata, visualize finite automata, and apply finite automata. This toolbox was developed mainly by Gertjan van Noord at the University of Groningen in cooperation with Dale Gerdemann.
- **W**eb-Interface for searching GermaNet (not yet published)
- **WebLicht** (Web-based Linguistic Chaining tool) – WebLicht is a web-based service environment for the integration and use of language resources and tools. The integrated web services are part of a prototypical infrastructure that was developed to facilitate chaining of LRT services.

  The WebLicht web application is developed jointly by the Universities of Leipzig, Stuttgart, and Tübingen. Its main parts features are a Service Oriented Architecture, a Repository, a user interface are developed individually at the different institutes. Individual services are contributed by the named partners and by the BBAW. The internal data exchange format was developed as part of the D-SPIN-work on WebLicht and is called “Text Corpus Format” (TCF).

3.3 Institut für Deutsche Sprache (IDS) Mannheim

The Institute for the German Language in Mannheim hosts a large amount of language resources and tools that are maintained and continuously further developed by own resources and adapted to the CLARIN/D-SPIN-infrastructure by own and D-SPIN-resources.

Written corpora

- **German Reference Corpus (DeReKo)** – largest collection of electronic corpora of contemporary written German (currently 3.75 billion words)

- **Historical corpora (70 million words)**

Archives of speech corpora

- **Archive of Spoken German (AGD)** – recordings of dialects, conversations, institutional interaction, mediated talk (39 spoken corpora: 16300 audio, 900 video recordings; approx. 4400 hours; 6650 transcripts)
- **Database of Spoken German (DGD)** – searchable online-database of the AGD: 6 spoken corpora: 3900 transcripts, 1030 sound-aligned transcripts, approx. 2100 hours, partially aligned; additional 15 spoken corpora

Lexical resources

- **Elexiko** – dictionary of contemporary German
- Dictionary of neologisms and lexical innovations

Tools

- **COSMAS-II** – second generation corpus search and analysis system
- IDS-analyzer for higher-order collocations
- IDS-lemmatizer
- **Morphisto** – morphological analyzer
- **Dupecheck** – near-Duplicate-Detection for large corpus archives
- **VICOMTE** – collocation explorer

Information systems (resources + tools)

- **OWID** – lexicographic information system and interface to all lexica developed at the IDS
- **GAIS** – conversation analytical information system
- **grammis** – grammatical information system
- **BDP** – bibliography on German grammar
- **OBELEX** – online-bibliography for electronic lexicography
- **CCDB** – collocation database and corpus linguistic experimentation-platform

3.4 Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)

The D* project groups at BBAW, i.e. DWDS, DTA and DLEX, have compiled various lexical resources and tools, partially by their own means and partially with third-party funds.

Overview of the language resources at BBAW

The following resources are accessible via an integrated portal at http://www.dwds.de. A substantially revised and extended version will be available soon and can already be accessed through the URL http://beta.dwds.de.

Lexical Resources

- **Wörterbuch der deutschen Gegenwartssprache** (in short: WdG, Klappenbach/Steinitz 1952-1977) – The WdG is a six-volume dictionary of the German language. It has been compiled at the Academy of Sciences in Berlin between 1952 and 1977. Since 2003 the text of this dictionary is available in digital form. This digital edition has been annotated in conformance with the TEI guidelines for printed dictionaries. Between 2007 and 2009 a deeper structural analysis and annotation was performed. As a result, much more of the information fields are annotated with semantically more adequate labels. The WdG comprises of full articles for 90,000 headwords and reduced articles for another 30,000 compound headwords.
- **Etymologisches Wörterbuch des Deutschen** (in short: EtymWB, Pfeifer 1993) – This dictionary, which has first been published in 1989 and has seen three editions in print has been converted into a digital document in the recent years. The complex article structure has been partially analysed and annotated in conformance with the TEI guidelines for printed dictionaries. This dictionary comprises of 22,000 headwords.
- **DWDS-Wörterbuch** (in short: DWDS-WB) – This dictionary is a general language dictionary of the German language. It has been compiled from the substance of the WdG (see above). This dictionary is currently maintained and extended. It is planned to be a substantial revision of the WdG. One of the major features of this dictionary is the in-depth and careful treatment of orthographic variants. Explicit reference is made to the status and origin of these variants. A second major feature is the information on the pronunciation of the headwords. All headwords are being spoken by a professional speaker and carefully checked by experts in the field of German phonetics. These pronunciations are accessible as audio files. Substantial extensions of the number of articles are planned for the years after 2012.
- **Wortwarte** A collection of neologism, maintained for ten years now and constantly growing. This resource currently contains 32 000 entries with morphosyntactic descriptions and one citation at least. 1000 entries additionally consist of a definition.

Corpora

All corpora are annotated in conformance with the TEi guidelines (version P5).

- The **DWDS corpus** is a balanced reference corpus of the German language of the 20th century of roughly 100 million tokens. It is balanced with regard to the distribution of texts over text genres and the 10 decades of this century. Recently we were able to close some gaps (portions of some genres in some decades had been missing). We are currently working on the compilation of a comparable corpus for the first decade of the 21st century. We could already make available some fictional as well as non-fictional texts of this decade.
- The **DWDS complementary corpus** has been compiled in the context of a project on collocations and their lexical description ("Kollokationen im Wörterbuch"). It contains roughly 1 billion tokens of newspaper text, the largest part of it from the late decades of the 20th century. The corpus will be updated continuously and we are currently acquiring new texts from the earlier decades of the century.

Additionally, we can offer some special corpora:

- **"Juilland-D"** – A compilation of German texts in accordance with Juilland's method of corpus compilation.
- A **spoken language corpus** – transcripts of speech from all decades of the 20th century. This corpus contains roughly 2.5 mill tokens.
- "**DDR-Korpus**" – Documents from the former GDR (1949-1999) and fictional text written by authors in the GDR, but published, for political reasons, in Western Germany.
- "**Korpus jüdischer Periodika**" ("Compactmemory") – The corpus of Jewish periodicals – written in German - of the period between 1887 and 1938 contains roughly 27 million tokens. All texts are linked to the images of the pages of these periodicals.
- **Newspaper corpora** – a.o. Die Zeit, tagesspiegel, Berliner Zeitung. There are, however, access restrictions on some of these corpora.
- **Deutsches Textarchiv** (DTA) - This project aims at digitizing influential texts of the period from 1650 to 1900. As such, it is a good complement for the 20th century corpora of the DWDS project. The data, texts as well as the corresponding images of the book pages, are made accessible to the public via the project webpage (www.deutschestextarchiv.de). In addition to making the texts and images available, an annotation tool is offered with which researchers can add their own individual annotations to these texts ("cumulative work on text corpora").

Tools

The following tools have been developed by BBAW. Some of them can be accessed via web services:

- **DWDS/Dialing Concordancer** (DDC) – An Indexing and Concordancing tool, which allows for complex queries on linguistically annotated texts. This tool is currently used to access our corpora as well as access tool the C4 corpus. DWDS-Tokenizer. This tool is available through a web service.
- **POS-tagger "moot"** – The tool uses statistical methods for the disambiguation of the POS-tags. In addition to the classical bigram/trigram-disambiguation strategy it uses handcrafted sets of potential analyses (so-called lexical classes) for each entry word. These rules constrain the potential and contextually driven analyses to such analyses which are plausible given the morphological structure of the word. Compared to a simple HMM model this extension leads to an error reduction of 21 %. The tagger is accessible through a web service.
- **"Syntactic Constraint Parser"** (SynCoP) - SynCoP is a system for the automatic syntactic analysis of corpora. The system has been implemented by Jörg Didakowski and uses a finite-state approach. The linguistic approach is dependency parsing. Jörg Didakowski will further develop this tool for the D-SPIN project. It will be possible to link some special-purpose tools, e.g. for named-entity recognition to the output of the parser.
- **Named-entity recognizer** – This tool uses the analyses of the parser and handcrafted rules to detect and mark named entities of several classes (e.g. names of persons and locations). The

tool will be made available through a web service. This tool is designed for the analysis of German text of earlier periods, in particular 17 – 19 century.

- **Statistical word profile generator** – This tool, which uses techniques comparable to those of Killgarriff's „word sketch engine", extracts and presents pairs and triples of words which co-occur with statistical significance and within qualified syntactic constructions (i.e. "colligations").
- **Word frequency timelines** – Based on the DWDS-Kernkorpus and its metadata, the frequency of occurrence of a word and its distribution over the text genres and decades are recorded and visualized.

3.5 University of Leipzig

The NLP group of the department of computer science of the University of Leipzig offers LRT resources and tools that may be used for free in teaching and research projects. During the D-SPIN project, some of these tools and resources could be improved using funds provided by the department.

Overview of resources and tools developed and maintained in Leipzig

- **Project "Deutscher Wortschatz":** For many years the project "Deutscher Wortschatz" offers large corpora and statistical data on the "Corpus-Portal" http://corpora.informatik.uni-leipzig.de. The statistical analysis of these corpora is based on language independent, automatic and semiautomatic algorithms. Currently these corpora are available in 57 different languages. The computed data consist of:
  - Frequency and frequency-class
  - Sentence based co-occurrences
  - Right and left neighbors
  - Example sentences (including source reference)

German corpora additionally contain the following data:

  - Base form
  - Grammatical data
  - Subject area
  - Synonyms
  - Morphology

- **ASV-toolbox** – The ASV-toolbox represents a collection of tools that allow the analysis of texts. The collection consists of algorithms for POS-tagging, base form reduction, named entity recognition, terminology extraction and several others. Most of them are optimized and tested for usage on huge amounts of data. The ASV-toolbox is intended to be used in order to demonstrate the contained NLP-algorithms in teaching lessons, but may also be used freely in scientific projects.

- **Web services** – Since 2004, the NLP group offers SOAP based web services that allow easy access to several resources and the usage of algorithms developed in Leipzig. These web services are currently used on a large scale in several research projects and by many researchers all over the world. Between September 2006 and July 2009 nearly 160 million requests were computed. The offered web services consist of:
    - Access to data of the project "Deutscher Wortschatz" (frequency, example sentences, left and right neighbors, co-occurrences, base forms, ...)
    - Co-occurrence statistics
    - Synonyms, computed on the basis of co-occurrences profiles of a word
    - A graph based clustering algorithm
    - A sentence segmenter

3.6 University of Frankfurt

The TITUS project goes back to the year 1987. Since then, digitized texts in old Indo-European languages have been collected and processed by Jost Gippert.

Since 1994, the project bears the name **TITUS** (TITUS = Thesaurus Indogermanischer Text- und Sprachmaterialien = Thesaurus of Indo-European Text and Language Materials) and is hosted by the Institute for Comparative Linguistics at Frankfurt University. From the late nineties on, the texts have been available on the web site of the institute (http://titus.uni-frankfurt.de/texte/texte2.htm).

In the meantime, the coverage of the collection has been expanded considerably, in so far as both younger stages of languages and (increasingly) non-Indo-European (mostly Caucasian) languages have been included.

Several texts have a special preparation by the inclusion of parallel versions, metric or grammatical information. Moreover, the TITUS database includes the Old Turkic corpus of the VATEC project (Vorislamische Alttürkische Texte: Elektronisches Corpus) as well as the processed version of the audio-visual recordings from the projects ECLinG (Endangered Caucasian Languages in Georgia) and SSGG (The sociolinguistic situation of present-day Georgia). The whole content is registered in a database and retrievable with a flexible search engine.

Most texts are freely available.

On the whole, TITUS comprises a volume of about 30 GB of data. Available corpora are listed below:

- **Indic corpora**
- Vedic corpus
- Corpus of Classical and Epic Sanskrit
- Corpus of Buddhist Sanskrit
- Pali corpus
- Prakrit corpus
- Rajasthani corpus
- Hindi corpus
- Dhivehi corpus

- **Iranian corpora**
- Avestan corpus
- Old Persian corpus
- Khotanese Saka corpus
- Tumshuqese Saka corpus
- Sogdian corpus
- Parthian corpus
- Middle Persian corpus
- New Persian corpus
- SSGG corpus Laiji
- Ossetic corpus
- SSGG corpus Ossetic

- **Anatolian corpora**
- Hittite corpus
- Luvian corpus
- Palaian corpus
- Lydian corpus
- Corpus of Lykian und Milyan
- Pisidic corpus
- Carian corpus

- **Tocharian corpus**
- East Tocharian corpus
- West Tocharian corpus

- **Armenian corpora**
- Old Armenian corpus
- SSGG corpus Armenian

- **Baltic corpora**
- Old Prussian corpus
- Latvian corpus
- Lithuanian corpus

- **Slavonic corpora**
- Old Church Slavonic corpus
- Old Czech corpus
- Old Polish corpus
- Old Slovene corpus
- Old Croatian corpus
- Old Russian corpus
- SSGG corpus Russian
- Old Sorbian corpus

- **Germanic corpora**
- Gothic corpus
- Nordic corpus
- Old English corpus
- Old Frisian corpus
- Old Saxon corpus
- Middle Low German corpus
- Old Dutch corpus
- Old High German corpus
- Middle High German corpus
- Corpus of Early New High German and Dialects

- **Greek corpora**
- Mycenaean corpus
- Corpus of Homeric Greek
- Corpus of Classical Greek

- **Italic corpora**
- Oscan corpus
- Umbrian corpus
- Latin corpus
- Old French corpus
- Old Italian corpus
- Old Portuguese corpus

- **Celtic corpora**
- Old Irish corpus
- Middle Welsh corpus

- **Other Indo-European corpora**
- Albanian corpus
- Phrygian corpus
- Corpus of fragmentarily attested languages

- **Caucasian corpora**
- Old Georgian corpus
- Middle Georgian corpus
- Modern Georgian corpus
- SSGG corpus Georgian Dialects
- Laz corpus
- SSGG corpus Laz
- Megrelian corpus
- SSGG corpus Megrelian
- Svan corpus
- SSGG corpus Svan
- ECLinG corpus Svan
- Batsbi corpus
- SSGG corpus Batsbi
- ECLinG corpus Batbsi
- Udi corpus
- SSGG corpus Udi
- ECLinG corpus Udi

- **Turkic corpora**
- VATEC corpus Old Turkic
- SSGG corpus Azerbaijanian

- **Semitic corpora**
- Hebrew corpus
- Arabic corpus
- Syriac corpus

- **North Picene corpus**

- **Proto-Cretan corpus**

3.7 Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) Saarbrücken

Lexica

- **IDX** (under License for commercial use) – the underlying lexicon of the IDX system, which is a professional text indexing system with high-quality linguistic knowledge. It achieves an excellent level of consistency in the indexing of large amounts of German or English text. IDX is able to determine the base words of derivations and compounds. IDX is also available for Italian.
- **Mmorph Lexicons** – MMorph originally has been a result of the Multext Project (http://aune.lpl.univ-aix.fr/projects/multext/). While the software component only underwent minor modifications, the linguistic resources have been substantially evolved since, and are maintained at DFKI. The lexicon covers English (approx. 200,000 entries), German (830,000), French (225,000), Italian (330,000), and Spanish (570,000)

Corpora

- The **Tiger Corpus /Treebank** (developed by the University of Saarland, IMS in Stuttgart and University of Potsdam). The TIGER Treebank (Version 2.1) consists of app. 900,000 tokens (50,000 sentences) of German newspaper text, taken from the Frankfurter Rundschau. The corpus was semi-automatically POS-tagged and annotated with syntactic structure. Moreover, it contains morphological and lemma information for terminal nodes.

Tools

- **SProUT (Shallow Processing with Unification and Typed Feature Structures)** – SProUT is a system for partial analysis of texts. It is used in particular – though not exclusively – for named entity recognition (NER) and opinion mining. The simple recognition of named entities includes, among others, persons, locations, date and currency expressions, functions, companies and organizations.

  SProUT currently recognizes named entity expressions in German, English, French, Spanish, Chinese and Japanese with high quality. Linguistic resources for other languages are integrated on a continuous basis.

  SProUT is implemented in Java and C and is equipped with a Java API. The system can thus be easily integrated. SProUT processes text files and delivers structured results in XML format. For grammar developers, SProUT offers a development and test platform with a comfortable graphical user interface. This way, language resources can be adapted to individual requirements.
- **Heart of Gold (Middleware Architecture for the Integration of Deep and Shallow Natural Language Processing Components)** – Heart of Gold is middleware architecture for the integration of deep and shallow natural language processing components. It provides a uniform and flexible infrastructure for building applications that use Robust Minimal

Recursion Semantics (RMRS) and/or general XML standoff annotation produced online by natural language processing components.

The main purpose Heart of Gold was developed for is tight integration of various shallow natural language processors with a deep parser.

The aim of the integration is to increase robustness of deep grammars for various languages such as English, German, Japanese, Greek and Norwegian. Deep grammars can be developed with the Linguistic Knowledge Builder LKB, compiled to a binary grammar image, and run within Heart of Gold.

Although the focus of Heart of Gold is deep-shallow integration, the framework itself is generic and hence can also be used to annotate corpora automatically and multi-dimensionally, combine multiple purely shallow systems on XML basis, or to integrate other deep parsers.

3.8 University of Stuttgart

At IMS Stuttgart, resource production and adaptation is carried out at several levels:

- Preparation of corpus data: provision of an automatically annotated fragment of the German Web-as-Corpus
- Enhancement of lexical resources for tools: tagger lexicons for TreeTagger and improved morphological analysis rules for the German morphology system SMOR
- Work on the definition of the Text Corpus Format (TCF) used in the D-SPIN web services, as implemented in the WebLicht tool
- Development of evaluation methods and Gold Standard data for the evaluation of (German) morphology systems, in connection with the evaluation of the STR system SMOR.

These activities have been closely coordinated with the D-SPIN partners. A large part of the actual work was financed from IMS's own funding, as part of the "Grundausstattung" (Prof. Hinrich Schütze, Helmut Schmid, Ulrich Heid, Edgar Hoch and student assistants: management of the data production work, TCF development, SMOR evaluation).

Corpus Data

At IMS, the raw version of Baroni and Kilgariff's fragment of the German contributions to the internet (DeWaC) has been turned into a processable corpus. To this end, non-sentence parts (a considerable chunk of data downloaded by the original authors was just arbitrary word lists) had to be removed from the future corpus; furthermore, duplicate sentences were also removed. To able to keep track of the source of the remaining data, a version of the material has been encoded as a corpus for the corpus query tool CorpusWorkBench (CWB), which contains the URL of the source of the text as metadata. Due to the heterogeneity of the DeWaC data, more fine-grained metadata annotations seem unfortunately hardly possible.

Furthermore, an objective of the work was to provide the subset of sentences from the **German web as corpus (DeWaC)** which can be syntactically analyzed by the dependency parser FSPAR: the

objective of this work is to provide a sizeable web-corpus of mostly well-formed sentences which can be used as a basis for lexical analysis in the future, e.g. frequency data for lexical items, material on collocations, syntactic constructions etc.; the resulting „cleaned“ corpus contains ca. 950 M words. For example, the extraction of candidate items for work on the evaluation and improvement of the morphology system SMOR is based on this corpus.

The provision of the parseable fragment of DEWaC is planned for the summer of 2010; work is still ongoing, but raw versions of the corpus are already available locally.

Enhancement of lexicons for NLP tools

Both TreeTagger and SMOR have been provided as web services, in the framework of D-SPIN. Over the first two years of the project, both resources have been constantly updated and enhanced, and improved versions have been made available via the WebLicht tool in the course of the project.

TreeTagger is a stochastic part-of-speech tagger developed by Helmut Schmid in 1994. It uses an internal lexicon containing word forms, the pertaining lemmas, as well as part-of-speech tags according to the Stuttgart/Tübingen Tag Set (STTS). The tagger has been trained on a manually annotated corpus of roughly 200,000 running word forms.

The tagger lexicon has been constantly enlarged, for example by adding lexical items from different domains of specialization, when German specialized language texts have been dealt with. The procedure to do so relies on the fact that TreeTagger can signal items not contained in its lexicon, when annotating texts. Then, lists of such “unknown” items have been semi-automatically annotated for lemma and part-of-speech, using the morphology system SMOR for the production of lemma and POS hypotheses. Manual correction and verification was used to ensure quality. The resulting additional material has been integrated into TreeTagger’s lexicon. This improves TreeTagger’s performance on arbitrary texts.

It is planned to carry out a similar updating exercise on the German Web-as-Corpus, as soon as the cleaned-up version (see above) is available.

The German morphology system **SMOR** is a finite-state-based system for both inflectional and word formation analysis (and generation) of German. SMOR relies on a substantial lexicon of nouns, verbs and adjectives, their sets, inflectional classes, morphologically relevant features, etc., as well as on a set of rules covering inflectional classes, as well as processes of prefixing, suffixing, conversion and compounding. SMOR has been made available as a web service within the WebLicht-Tool.

Improvements of SMOR on the one hand concern its lexicon and rule set, on the other hand the elaboration of a methodology to evaluate morphology systems at large, which will be applied to SMOR in the first place. The contact of the enhancement of lexicons and rules, erroneous analysis observed in regular tests, have been traced back to lacking or incomplete lexical information and/or to rules which over-generate or lead to incorrect analysis in another way. As a consequence, lexical knowledge has been added and numerous rules have been refined (Helmut Schmid, Ulrich Heid).

In the framework of the development of evaluation methods for morphology systems, a corpus based list of manually corrected morphological analysis from the Web-as-Corpus data discussed above is currently being prepared as a Gold Standard. This list will contain randomly selected German word forms of medium to top frequency, as well as their inflectional analysis, their word formation

analysis, as well as their inflectional paradigm, where necessary. This material will be made available to the research community as a test suite for morphology systems. It will comprise, in its first version, roughly 1,000 word forms and all their analysis at both levels indicated above. We expect this list to be available, along with results from SMOR and the pertaining updates of the SMOR morphology, in the late summer of 2010. Alongside, we plan to provide two separate morphology web services: One only providing inflectional analyses, and one leading to a morpheme decomposition of complex word forms: this distribution seems to be closer to the needs of different sets of users than a single service just providing a full morpheme decomposition.

IMS has also made available BitPar as a tool in the WebLicht chain. BitPar has considerably been improved, outside D-SPIN (Grundausstattung: Dr. H. Schmid, as well as SFB-732), and the improved version is included in the D-SPIN tool chain.

Finally, as part of IMS's complex web service for collocation extraction, the techniques used for multi-parametric collocation identification and classification have been reworked to (i) run on parsed data (FSPAR), (ii) extract a wide range of morphosyntactic features, (iii) include tools for corpus comparison. This work was completely carried out on IMS's funding (Grundausstattung), as only the wrapping for a web service was paid from D-SPIN funds (M. Kisselew).

3.9 University of Gießen

Gießen is not part of the work package 5 and hence does not implement any resources in the framework of the D-SPIN project.

4. Implementation activities

The standardization and integration of German resources is one of the core activities of D-SPIN WP5. D-SPIN partners who provide a multitude of LRT of different types, have undergone curation, improvement, standardization and more on their language resources and tools.

4.1 Max Planck Institute (MPI) Nijmegen

Activities concerning development and adaption of LRT

- All archived resources are already based on international standards and they are associated with metadata descriptions. The metadata descriptions include persistent identifiers that uniquely identify the instances of the stored objects. Versioning methods are in place. So with respect to the language resources no special activity will be required except for regularly checking the consistency of the archive and improve the metadata quality. To the whole archive the Data Seal of Approval quality assessment procedure is applied.
- Within CLARIN/D-SPIN a number of developments are currently being carried out:
    - Several tools will be equipped with APIs so that they can be invoked by other services. Some already have APIs, but they are mostly not well documented. In particular the TROVA search engine will get an API so that searches through the whole archive can be invoked from the outside. This is important for the demonstrator project.
    - The **Virtual Language Observatory** will be continuously improved to map the various metadata sets that are harvested. The quality is very heterogeneous, which requires much polishing and screening which to a large extent has to be carried out manually or with the help of scripts. It is meant to be the portal technology for all CLARIN language resources and tools. An improved interaction between the GIS viewer, the Facetted Browser and the Catalogue is being intended.
    - The **CMDI (CLARIN component based metadata infrastructure)** is the new flexible metadata format, which has been worked out in the first two years. Various infrastructure components are currently being worked on by various CLARIN/D-SPIN partners with the intention to offer a working infrastructure in 2010.
    - The **ARBIL metadata editor and organizer** have been launched in 09 and have been tested already. Currently the functionality is being extended to serve as CMDI editor and to replace the old IMDI infrastructure tools completely. The CMDI functionality should become available in 2010 as well.
    - **REPLIX** is a joint project between CLARIN and DEISA to develop and test safe data replication, which will be very important not only for CLARIN. This replication is based on persistent identifiers and information about the instances of each resource. Currently, tests are carried out to use iRODS for this task that is so crucial for long-term preservation and data authenticity.
    - **ISOcat** is a software tool to register and edit data categories (concepts), which is based on ISO 12620. It is already in use and has been widely debugged. However, still some essential functionality such as database replication is required to be added.

Planned activities

- Documentation and implementation of APIs (WSDL-SOAP and/or REST) for many of the tools.
- Extension and improvement of the Virtual Language Observatory, the basic machinery, is in place.
- Development of the new CMDI infrastructure for component based metadata descriptions.
- Extension of ARBIL to become the editor for CMDI metadata.
- Testing out methods for safe data replication in research infrastructures.
- Finishing the ISOcat developments.

4.2 University of Tübingen

Activities concerning development and adaptation of LRT:

- The WebLicht tool chain has been constructed (see above).
- Implementation of an AJAX-driven web application as user-friendly interface for WebLicht
- Addition of the OpenNLP tool chain into WebLicht
- Integration of the centralized repository database (developed in Leipzig) into WebLicht
- Dissemination of the WebLicht platform in the linguistic community; publication of a detailed tutorial how to participate in WebLicht; applications of Finnish colleagues are already integrated
- GermaNet has been transferred into an XML database. Before, GermaNet consisted of single text files in a "lexicographer's files" format. They were converted into an XML-synset format and transferred into a database.
- Integration of GermaNet into WebLicht

Planned activities:

- Integration of TüBa-D/Z into WebLicht
- Extension of TüBa-D/Z for the layer of lemmata. As the lemmatization will be done manually, research on collocations and quantitative analyses will be facilitated.
- Parameterization of the individual web services of WebLicht
- Addition of further complete tool chains into WebLicht (NLTK, etc.)
- Implementation of "superservices", services that combine several individual web services into complete processing chains

4.3 Institut für deutsche Sprache (IDS) Mannheim

Activities concerning development and adaption of LRT

Virtually all resources listed above were further developed or extended and new ones were created during the reported period. For example:

- The German Reference Corpus DeReKo was expanded by approx. 750 million words
- DeReKo was endowed with three concurring part-of-speech annotation layers
- The corpus FOLK with systematically collected spoken German was established at the AGD and is being further developed
- A corpus of call-center-communication and a corpus for migrational linguistics were newly established

Directly related to an adaption to the CLARIN/D-SPIN infrastructure were the following activities:

- COSMAS-II was endowed with a Shibboleth interface and now acts as a Service-Provider in the CLARIN prototype federation and AAI
- COSMAS-II was extended by prototypical REST and SOAP web services for accessing DeReKo and other corpora at the IDS.
- An initial integration of the created web services into WebLicht was prepared.
- The versioned metadata-database for DeReKo was extended and made harvestable via the OAI Protocol for Metadata Harvesting (OAI-PMH).
- A first prototype of a registry for virtual collections was implemented. Virtual collections are collections of language resources that can be physically distributed over many LRT-centers. One goal is for example, to be able to define a persistent virtual corpus out of all corpora in a network of centers that is representative with respect to a specific research question and a specific basic population, by defining properties of the contained texts.

Planned activities

- Extend the web service interface of COSMAS-II to make more of its functionality accessible in D-SPIN/CLARIN, particularly via WebLicht
- Continue work on virtual-collections (GUI for defining collections, integration in CMDI, …)
- Make metadata for speech corpora harvestable within the CLARIN framework
- Upgrade license agreements with some of the copyright holders of texts in DeReKo to allow for more unrestricted access on sentences in random order (feasibility study)
- Mapping of the IDS-XCES-format of DeReKo to TEI-P5

4.4 Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) Berlin

Recent processing steps

The schemata and the headword lists of the dictionary have been unified and aligned to facilitate the macro structural and micro structural access to all dictionaries and the presentation of bits of data from all resources.

Some of our NLP tools have been integrated into the WebLicht tool chain of web services. Input and output formats have been adjusted for this purpose.

Further activities

Drafting of a standardized data format for various lexical resources to be used within the WEBLICHT architecture.

Integration of the WdG (see above) as a semasiological lexical resource into WebLicht.

Link the C4-Corpus of German regional variants, its metadata and its search interface DDC with Weblicht in order to create a use case for the CLARIN Demonstrator

4.5 University of Leipzig

Recent activities

- Construction of a high quality French corpus (http://wortschatz.uni-leipzig.de/ws_fra/)
- Transformation of all data of the Wortschatz project to UTF-8
- First tests of services provided by the D-SPIN prototype in order to add additional annotations to the data of the "Deutscher Wortschatz"-project (for example POS-annotation based on the Tree-Tagger provided by the D-SPIN partners in Stuttgart)
- Participation in the definition of a standardized format that allows the usage of text based data in the D-SPIN prototype (D-SPIN Text Corpus and Lexicon-format)
- Implementation of REST-based web services compatible to the D-SPIN prototype. These web services provide access to data of the „Deutscher Wortschatz"-project and to several tools available at the department (sentence segmentation, tokenization)
- Implementation of a TopicMap-based tool (http://cls.informatik.uni-leipzig.de/), the „Korpusstatistikbrowser", that allows the statistical analysis of resources provided through the corpus portal. The tool was implemented in a joint effort of the NLP-group and the TopicMapsLab (http://www.topicmapslab.de/)

Planned activities

- Implementation of additional, D-SPIN prototype compatible web services, providing access to data of the „Deutscher Wortschatz"-project in many more languages
- Addition of some of the resources and algorithms provided by the eAQUA project (http://www.eaqua.net/) to the D-SPIN prototype

4.6 University of Frankfurt

Processing activities

As part of the D-SPIN project, the individual corpora are being converted into a D-SPIN compliant XML structure.

Main issues in the conversion are

- The consistent, standard compliant encoding of the languages involved in the form of an XML attribute for each word form
- The preservation of the structure both text inherent and edition or manuscript specific (i.e. integration of differing, non-congruent hierarchies in XML)
- The consistent stylage of TEI-compliant headers
- The retention of additional information (metrics, grammar, lexicon)
- The linking with parallel versions (e. g. the differing manuscripts of the Nibelungenlied, Latin texts with parallel Old High German translation or bible translations in different languages)

In the first phase, a conversion program was developed, which created XML files from the corresponding HTML versions of the texts. The first two issues mentioned before were already integrated. A number of texts of minor or medium size have already been converted.

XML files generated this way for Old High German and Old Saxon texts already have been used as the basis for further enrichment with grammatical and lexical information within the DFG project "Referenzkorpus Altdeutsch".

Problems remained in converting very extensive texts (the Rigveda consists of about 10 MB of pure text data). So the structure of the program had to be changed. Furthermore, the program had to be adapted to texts that are affected by the last two issues.

In the second phase of the project, the program has been modified and extended accordingly. The final version is imminent.

Planned activities within D-SPIN

- As soon as the conversion of all types of texts is available, the integration into WebLicht can be carried out.
- The integration of references to occurrences in parallel texts will need an intensive coordination with WebLicht.
- Furthermore, a comparison of the text retrieval on WebLicht with the possibilities of the TITUS search engine will be necessary. Some mechanisms of retrieval may have to be added or modified.

4.7 Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) Saarbrücken

Activities concerning development and adaption of LRT

- DFKI involved in many ISO TC37/SC4 Standardization activities, dealing with linguistic annotation, towards interoperability of linguistic resources.
- DFKI developed a tool for the semantic annotation of Fairy Tales (to be integrated in WebLicht).
- DFKI made one of its in house parser compliant to the on-going standardization work in ISO.

Planned activities

- DFKI will include some of its tools into WebLicht (NE, Dependency, Semantic Annotation of Fairy Tales), also in the context of the CLARIN Demonstrator

4.8 University of Stuttgart

Formats for Corpus Data Encoding

In the framework of the D-SPIN web services and especially of web service chaining, as carried out in the WebLicht tool, the need for D-SPIN-internal formats for text corpora annotated at different levels of linguistic description and annotation became apparent. In close collaboration with the University of Tübingen and other partners involved in the creation of the WebLicht tool chains, a processing-oriented internal format has been created which combines processing efficiency and easy convertibility towards upcoming international standards. This format, the D-SPIN text corpus format, TCF, is in the process of being compared with and mapped onto the upcoming ISO standards MAF, SynAF and LAF. D-SPIN has taken the role of an early adopter of these proposals for international standards, and coding examples from D-SPIN web service tools (such as H. Schmid's parser BitPar) have been provided for both the LAF meta model (July/August 2009) and the MAF/SynAF format proposals (February 2010). A close collaboration with the developers of these standards proposals (Prof. L. Romary, Nancy/Berlin and his group at HU Berlin) has been sought

(meetings in 09/2009, 02/2010, and planned for spring and summer 2010). Thereby, D-SPIN contributes to the verification, test and enhancement of the upcoming international standards proposals, as it attempts to use them in its tool chains as an exchange format.

The current WebLicht web services are based on TCF. Most of the TCF development at IMS was carried out by Helmut Schmid, Kerstin Eckart and Ulrich Heid (thus to a great part from "Grundausstattung"). TCF, its underlying reasoning and its relationship with the ISO proposals will be presented at the Linguistic Resources and Evaluation Conference (LREC) in May .

4.9 University of Gießen

Gießen is not part of the work package 5 and hence does not implement any resources in the framework of the D-SPIN project.

5. WebLicht – Web-based Linguistic Chaining Tool

The definition and implementation of web services is a second core theme of WP 5. The D-SPIN Partners in Stuttgart, Leipzig, and Tübingen jointly developed the web portal WebLicht (Web-Based Linguistic Chaining Tool, http://weblicht.sfs.uni-tuebingen.de/englisch/weblicht.shtml). They and the BBAW integrated their linguistic tools into WebLicht.

The integrated web services are part of a prototypical infrastructure that was developed to facilitate chaining of LRT services. WebLicht allows the integration and use of distributed web services with standardized APIs. As a first external partner, the University of Helsinki contributed a set of web services to create morphologically annotated text corpora in the Finnish language.

WebLicht is a Service Oriented Architecture, which means that distributed and independent services are combined together to a chain of LRT tools. A centralized database, the repository, stores technical and content-related metadata about each service. All services are registered in a central repository located in Leipzig. Also realized as a web service, it offers metadata and processing information about each registered tool, e.g. information about the creator, name and address of the service. Additionally, input and output specifications of each web service are stored. This information is used by the chaining algorithm to determine, which combinations of services form a valid processing chain. The chaining algorithm was developed in Leipzig, too, and is currently accessible through web services offered along the previously mentioned repository services. It is used by the WebLicht web interface during the orchestration process of a service chain.

The WebLicht web interface is developed and hosted in Tübingen. It provides an overview of the available web services and their chaining combinations.

Plain text input to the service chain can be specified in one of three ways. It can be typed in, uploaded in a file, and sample texts are offered.

The D-SPIN Text Corpus Format TCF, developed in Stuttgart, is used by WebLicht as an internal data exchange format. The TCF format allows the combination of the different linguistic annotations produced by the tool chain. It supports incremental enrichment of linguistic annotations at different levels of analysis in a common XML-based format. The TCF was designed to efficiently enable the seamless flow of data between the individual services of a Service Oriented Architecture.

The WebLicht platform in its current form moves the functionality of LRT tools from the users' desktop computers into the net. At this point, they must download the results of the chaining process and deal with them on their local machine again. In the future, an online workspace has to be implemented, so that annotated text corpora created with WebLicht can also be stored in and retrieved from the net.

6. Summary

D-SPIN members already have quite a lot of language resources and tools available that they developed to a large extent on-site. Thus, these LRT with rights to D-SPIN partners can be standardized, worked over and finally be published and integrated into new cooperations; of course depending on their legal restrictions. An overview of the relevant legal aspects is provided in the D-SPIN Report R7.1 "Legal Aspects in the Provision of Language Resources: The German Context".

Depending on the grouping of individual components into "resource-entities", the number of available language resources may vary widely. E.g., the MPI Nijmegen stores about 1 million files on 60,000 audio video sessions in their multimedia/multimodal speech archive. The sessions can be grouped differently depending on users' needs. Their most convenient grouping is into ca. 200 collections on ca. 200 languages.

The amount of language resources of the D-SPIN partners ranges from one to 200 with Tübingen, Stuttgart, IDS, DFKI having under 10 resources, the BBAW, Leipzig and Frankfurt naming 13 – 69 resources and the MPI with about 200 resources. In total, about 350 resources are listed.

As to tools, D-SPIN partners developed up to 16 applications, ranging in complexity from converters to entire interfaces, with the Virtual Language Observatory, ISOcat and WebLicht being probably the most prominent ones. In total, D-SPIN partners named almost 50 applications.

Ongoing work among the D-SPIN partners includes improvement and extension as well as adaptations of standards for LRT and harvesting of metadata. Also, several new corpora and tools are being developed.

The aim of WP 5 to define and implement web services and to integrate and connect resources is heading into the right direction. In WebLicht, currently there are about 60 web services online, with 50 being developed by the Universities of Stuttgart, Leipzig and Tübingen and by the BBAW. The University of Helsinki and RACAI (Research Institute for Artificial Academy, Romania) are the first international partners to integrate their resources into WebLicht; several interested parties are to follow.