D-SPIN

D-SPIN Report R1.2b: Progress Report of the Project and the various Working Packages – Year 2

April 2010

D-SPIN, BMBF-FKZ: 01UG0801A

Deliverable R1.2b: Progress Report of the Project and the various Working Packages – Year 2

Responsible: Erhard Hinrichs

Editors: Erhard Hinrichs, Kathrin Beck, Lothar Lemnitzer

Summary

D-SPIN is the German counterpart of the European Research Infrastructure project CLARIN (Common Language Resources and Technology Infrastructure; cf. http://www.clarin.eu). The ultimate objective of CLARIN is to create a European federation of digital repositories that include language-based data, to provide uniform access to the data, wherever it is, and to provide available language and speech technology tools as web services to retrieve, manipulate, enhance, explore and exploit the data. The primary target audiences are researchers in the humanities and in the social sciences, and the aim is to cover all languages relevant for the user community. The objective of the current CLARIN Preparatory Phase Project (2008 – 2010) is to lay the technical, linguistic, and organizational foundations, to provide and validate specifications for all aspects of the infrastructure (including standards, usage, intellectual property rights).

Similar goals are pursued by the D-SPIN project on the national level. Work in D-SPIN is carried out in close collaboration with CLARIN. Within the CLARIN federation, the focus of D-SPIN is on resources tools developed in Germany and on their integration through web services (WPs 2 & 5). Besides, D-SPIN is addressing potential users of the infrastructure through WP 3's case studies and requirements analysis, and WP 6's training activities, and covers legal and ethical aspects. (WP 7). D-SPIN is also actively cooperating with other national projects and initiatives (WP 4).

This report gives an overview of the project activities in the second year of the project, i.e. April 2009 – March 2010.

Introductory Remarks

1. D-SPIN is the German counterpart of the CLARIN project. We will therefore devote sufficient space to report on CLARIN activities whenever they are relevant to the D-SPIN project. On the CLARIN website, there is a list of CLARIN documents – in particular, minutes of key meetings and deliverables – which will enable readers to follow the discussions and decisions taken on the CLARIN level.
2. D-SPIN activities are supported by both the BMBF funding as well as by third party funding which has been dedicated explicitly to the CLARIN/D-SPIN activities. In this report, we will report on the D-SPIN activities without mentioning the source of funding for each activity. For details we refer to the intermediate reports ("Zwischenbericht") of each of the partners where the sources of funding are listed.
3. Activities are listed per work package, if not mentioned otherwise.
4. The purpose of this report is to provide a short overview of the activities carried out in the DSPIN project during the second year.
5. According to the schedule of deliverables, work packages 2 – 7 have produced more detailed reports about specific aspects of their work. These deliverables are entitled:
    - R 2.2 The German Resource Landscape and a Portal
    - R 3.2 + 4.1 Liaison Report and Roadmap
    - R 3.3 Case Studies – Intermediate report
    - R 5.2 Documentation of the D-SPIN preparation activities
    - R 6.1 Training materials and guidelines for their application
    - R 7.2 German Localization of CLARIN Model Agreements

WP1: Management and Dissemination

Work package 1 is devoted to the management of the project as well as to the dissemination of the results.

International Advisory Board

On 17th and 18th October 2009, a first meeting with the D-SPIN Advisory Board took place in Munich. Work of the individual D-SPIN partners as well as the current state of the work packages was presented. Especially the sustainability both of the LRT developed and of the knowledge obtained in the project are pointed out.

Members of the Scientific Advisory Board are:

- Helge Kahler (BMBF)
- Axel Horstmann (Volkswagen Stiftung)
- Christiane Fellbaum (Princeton University)
- Bernhard Neumair (GWDG, University of Göttingen)
- Neil Freistat (Maryland Institute for Technology in the Humanities)
- Paul Doorenbosch (Koninklijke Bibliotheek NL)
- Bente Maegaard (University of Copenhagen, CLARIN Liaison)

Meetings

The following plenary meetings were organized:

- A plenary meeting in Munich on 17th October 2009
- A plenary meeting in Stuttgart on 13th February 2010

Dissemination

The following public relations and dissemination activities have been initiated in this work package:

- The domain http://www.d-spin.org/ was reserved for D-SPIN. The D-SPIN web page is available via this address.
- D-SPIN organized a Language Resources Summit in Mannheim on 15th – 16th May 2009 with ca. 50 participating institutions. (For an overview see http://weblicht.sfs.uni-tuebingen.de/LRTSummit/Sprachressourcengipfel.html)

Liaison with CLARIN

There is a close and regular cooperation with the EU project CLARIN. Erhard Hinrichs, University of Tübingen and Peter Wittenburg, MPI Nijmegen are members of the CLARIN executive board and participate in the meetings of the CLARIN executive board.

D-SPIN is collaborating with CLARIN by holding joint workshops of mutual interest. During the second year, the following workshops were held:

- 27th – 29th April 2009, Tübingen: CLARIN/D-SPIN WP 2 workshop
- 15th – 16th May 2009, Mannheim: D-SPIN Language Resources Summit
- 4th – 6th June 2009, Freudenstadt: D-SPIN workshop on Finite State Technology und computational Morphology
- 11th July 2009, Berlin: D-SPIN/CLARIN/ISO meeting on ISO standards
- 28th September 2009, Berlin: D-SPIN workshop on Legal Constraints
- 28th September 2009, Berlin: D-SPIN/CLARIN/ISO meeting on ISO standards
- 19th – 29th November 2009, Leipzig: CLARIN WP 2 workshop

Project publications

- Ahlborn, Svetlana: *Mehrzweckorientierter eLearning-Kurs mit Einsatz rekursiver Übungen*: http://www.studiumdigitale.uni-frankfurt.de/events/nwt2009/index.html.
- Bankhardt, Christina: *D-SPIN – Eine Infrastruktur für Deutsche Sprachressourcen.* In Sprachreport Heft 1/2009 (25. Jg.).
- Broeder, Daan, Malte Dreyer, Marc Kemps-Snijders, Andreas Witt, Marc Kupietz, Peter Wittenburg: *Persistent and Unique Identifiers*. CLARIN-Deliverable D2.2.
- Eckart, Kerstin: *Repräsentation von Unterspezifikation in relationalen Datenbanken*, master thesis, Stuttgart: IPVS, 2009.
- Faaß, Gertrud, Ulrich Heid, Helmut Schmid: *Design and application of a Gold Standard for morphological analysis: SMOR in validation*. Poster presented at LREC 2010.
- Fritzinger, Fabienne et al.: *Werkzeuge zur Extraktion von signifikanten Wortpaaren als Web Service*. In: Wolfgang Hoeppner (pub.): *Akten des GSCL-Symposiums "Sprachtechnolgie und eHumanities"*, Duisburg, 2009.
- Gehrke, Ralf: *TITUS: datenbank- und internetorientierte Konzepte*. In: Hofmeister, Wernfried; Hofmeister-Winter, Andrea (pub.): *Wege zum Text. Überlegungen zur Verfügbarkeit mediävistischer Editionen im 21. Jahrhundert*. Grazer Kolloquium 17th – 19th September 2008. Tübingen: Niemeyer (Beihefte zu Editio, 30), S. 43–51, 2009.

- Heid, Ulrich, Fabienne Fritzinger, Erhard Hinrichs, Marie Hinrichs, Thomas Zastrow: *Term and collocation extraction by means of complex linguistic web services*. Paper published at LREC-2010.

- Heid, Ulrich, Helmut Schmid, Kerstin Eckart, Erhard Hinrichs: *A corpus representation format for linguistic web services: the D-SPIN Text Corpus Format and its relationship with ISO standards*. Poster published at LREC-2010.

- Erhard Hinrichs et al.: *WebLicht – Web-based LRT services for German*. Presentation and abstract for the workshop on Linguistic Processing Pipelines, GSCL-2009 (Potsdam).

- Erhard Hinrichs et al.: *WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure*. Poster published at LREC-2010.

- Lemnitzer, Lothar, Erhard Hinrichs/Andreas Witt: *Language Resources, Taxonomies and Metadat*a. Presentation and publication: Text Mining and Services, Leipzig, March 2009

WP 2: Technical Infrastructure

The work of WP 2 is closely related to the corresponding activities in CLARIN WP 2. This work is documented in detail in the following deliverables:

- D2R-1a Centres Network Formation (2009)
- D2R-1b Centres Network Formation - Centre types (2009)
- D2R-2a Federation Foundation - LRT (2009)
- D2R-2b Federation Foundation - PIDs (2010)
- D2R-3a Federation Building - v1 (2010)
- D2R-4 Registry Requirements (2009)
- D2R-5a Registry Infrastructure - v1 (2010)
- D2R-6a Web Services and Workflow Requirements - v1 (2009)
- D2R-6b Web Services and Workflow Requirements - v2 (2010)
- D2R-7a Web Services and Workflow Creation - v1 (2010)
- D2R-9a Cost Estimates - v1 (2010)

The deliverables are available on the CLARIN website and on http://www-sk.let.uu.nl/cgi/clarin_deliverables_list.ksh?wp.

In year two the focus was on the implementation of the requirement specifications that have been worked out in year one.

In this context, MPI worked out eight CLARIN short guides:

- Centre Criteria
- REPLIX
- Initial Service Provider Federation
- Standards for Text Encoding
- Create Metadata Now
- Virtual Collections
- Virtual Language Observatory
- Web Services Interoperability
- Web Services Infrastructure.

CLARIN Usage Scenarios

Participants of the Language Resources Summit in Mannheim on 15-16 May 2009 were invited to participate in CLARIN by registering their LRT in the Virtual Language Observatory of CLARIN, so that their metadata can be harvested and searched. A second offer was the participation of interested researchers in CLARIN work packages as external experts.

As a result of this presentation of CLARIN, several institutions became CLARIN members.

Furthermore the community was invited to create usage scenarios that will be considered in the next phase of D-SPIN/CLARIN. The two kinds of suggested projects were:

- Curation projects that integrate important data with a new infrastructure. Such projects

usually are quite short.

- Usage scenarios that bring together data and services and that help to realize new scientific projects.

D-SPIN received 17 usage scenarios and curation projects that are developed mostly in cooperation of several scientific institutes. They can bee seen at http://www.clarin.eu/suggestion-gathering-for-curation-projects-and-for-demo-cases.

CLARIN Centers

Requirements on CLARIN centers are refined and progress of the work of D-SPIN partners was discussed in Prague on 5th – 6th November 2009. The following D-SPIN members apply for CLARIN centers:

- MPI Nijmegen will be established as a service provider of infrastructure, resources and tools. The center focus will be on multimedia resources and on minority languages and special kinds of languages.
- IDS with their large collection of language resources will apply as resource provider. Corresponding tools will be made available as web service or web applications.
- University of Tübingen will participate as provider of tools, the German word net and several treebanks. In addition, with the development of WebLicht in cooperation with Universities of Stuttgart and Leipzig, it has a new direction of services and tools.
- BBAW is specialized in German lexica and corpora.
- DFKI is offering tools and metadata that can be accessed via the DFKI "Language Technology World" (http://www.lt-world.org)
- University of Leipzig will offer tools and statistical applications.

WP 3: Relations of the Project to Research in the Humanities

The University of Gießen is conducting a case study on using language resources for computer-assisted language learning. In the first step, a survey was conducted within different departments of the University of Gießen. The study was presented at the workshop "Cumulative work with corpora" at the BBAW on 27th March 2009. First results are:

- Teachers need to be trained on language resources, and they need easily accessible, efficiently usable resources in order to use them successfully in class.
- Learner corpora are useful both for language learners, for designing teaching material and for research in second language acquisition.
- Multimodal resources and sound archives can be used for teaching and learning pronunciation and intonation of foreign languages.
- Automatic feedback via an automated error analysis based on error-annotated learner corpora is one particular issue of ICALL.

The University of Frankfurt is establishing a liaison with the EuroBABEL project (http://www.esf.org/index.php?id=4632) and presented D-SPIN on their launch conference in Berlin on 11th – 13th September 2009. A case study on language resources in the context of documenting endangered languages is being developed.

The BBAW designed a questionnaire on the usage of language resources. It was published on the mailing lists "HsozKult" and "Gesprächsforschung". The quantitative analysis of the answers has been presented on the D-SPIN Advisory Board meeting in Munich in October 2009. A follow-up questionnaire for expert interviews is being developed.

WP 4: Liaison with other National Projects and Initiatives

On the level of the CLARIN project, the deliverable D5C-1 Survey of liaisons with other European projects and initiatives from March 2009 summarizes the direction of liaison activities of CLARIN with other projects and activities.

D-SPIN has liaison activities, among others, with the following groups:

- MPI and IDS are partners of the TextGrid consortium.
- MPI had several interactions with the DFN-Verein for integrating German institutions into the AAI and for signing the contract with the German Identity Federation
- DFKI is an active member in ISO TC37/SC4. BBAW and the Universities of Stuttgart and Tübingen are collaborating with TC37/SC4 in the development of standards for the syntactic annotation framework SynAF.
- The German coordination initiative on eHumanities under the aegis of the Volkswagen Stiftung, described in R1.2a, is still active.
- MPI and DEISA developed the REPLIX project
- IDS is member of the German competence network for digital preservation "nestor" (http://www.langzeitarchivierung.de/).
- University of Frankfurt, MPI and University of Chicago implemented the project "Rendering Endangered Languages Lexicons Interoperable Through Standards Harmonization".
- D-SPIN partners on various events presented the D-SPIN project, e.g. on the GSCL conference in Potsdam, September 2009 and the NEERI conference in Helsinki, October 2009.

WP 5: Language Resources and Tools

D-SPIN Report R5.1 "Documentation of the D-SPIN preparation activities" provides a detailed overview of the extensive work of standardization and integration of LRT that D-SPIN partners did on their language resources and tools.

Curation

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 of various kinds and sizes are provided by the D-SPIN members.

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.

WebLicht

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 funded and adopted, so that annotated text corpora created with WebLicht can also be stored in and retrieved from the net.

WP 6: Training and Education

In the second year of the D-SPIN project, two university courses for potential LRT-users (students of linguistics, students pursuing teaching certification) were developed and conducted at the Universities of Gießen and Frankfurt.

To support these courses, e-learning materials were developed at both universities. However, e-learning played different roles in the two courses. In Frankfurt, the course was based on self-created e-learning materials, which were specifically tailored for this course. In contrast, the Giessen course utilized existing textbooks and additional reading materials as well as student presentations. Except for course management and communications, e-learning played a rather supplementary role here.

Another significant effort of WP 6 in the second year of D-SPIN was to design and prepare the D-SPIN Summer School 2010 "Language Resources for Humanities".

- Focus of the summer school is on basic courses and on practical tutorials for resources and methods.
- The summer school will take place in the well-situated "Forschungskolleg Humanwissenschaften" in Bad Homburg from 30th August to 3rd September 2010.
- Participants will be 35 students of both D-SPIN member universities and other institutions.
- Students will select their preferred courses out of 12 classes that will be held mostly by D-SPIN members. External experts are invited for plenary talks and some courses, e.g. from TextGrid.

WP 7: Legal and Ethical Issues

The goal of this work package is to analyze the legal and ethical situation with respect to the compilation, the distribution and the usage of language resources, to collect and to analyze samples of usage licenses, and based on this analysis to give best practice recommendations to language resource providers from the legal and ethical point of view. The following activities have been initiated in the second year of the project:

- Localization of agreement templates developed at the CLARIN-EU-level, i.e. translating and adapting them to the German legal conditions
- Creation of model contracts for licenses between different parties/roles
- Examination of existing contracts for the usage of commercial annotation tools and for annotations in distributed infrastructures
- Collection of EULAs, license contracts and release forms
- Contribution to prototypical upgrade-contracts that extend license agreements between holders of rights and content providers