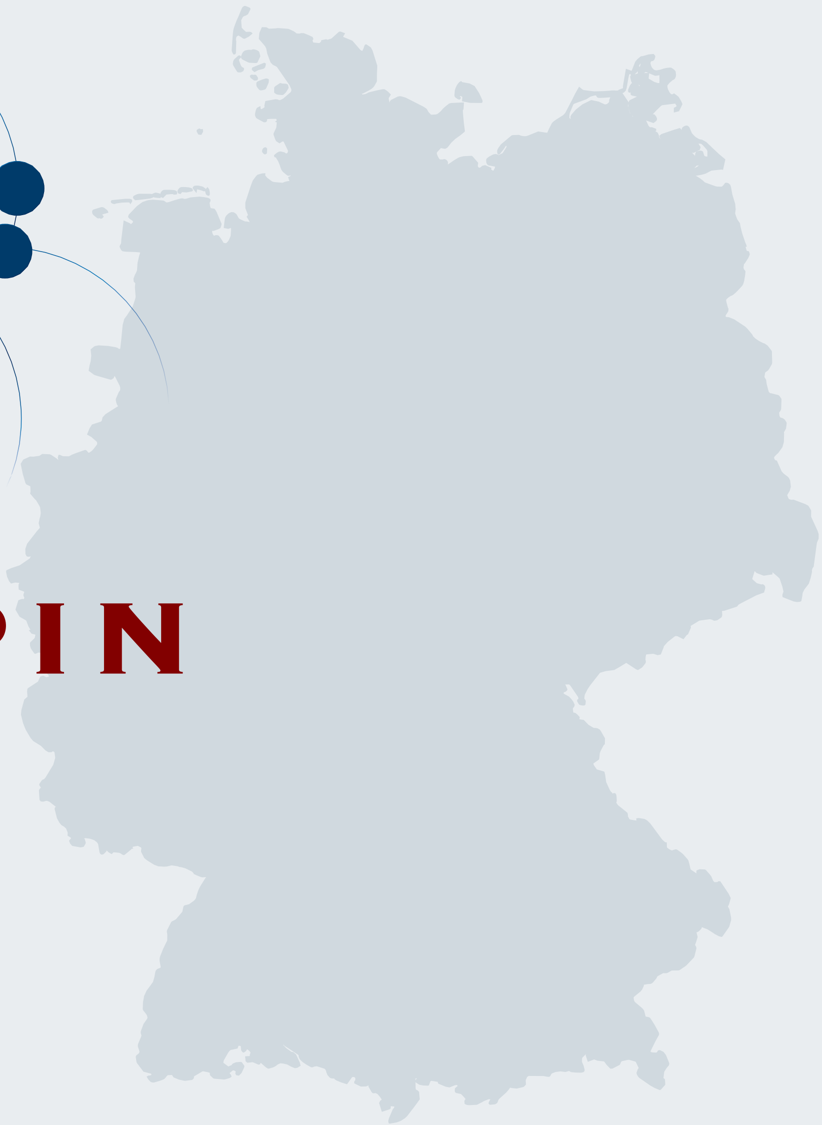

D-SPIN Report R3.1: Initial Requirements Analysis

June 2009

D-SPIN, BMBF-FKZ: 01UG0801D

Deliverable: R3.1:
Initial Requirements Analysis

Editor: Axel Herold, Alexander Geyken, Lothar Lemnitzer

Contents

Summary

D-SPIN is the German counterpart of the European Research Infrastructure project CLARIN (Common Language Resources and Technology Infrastructure, http://www.clarin.eu/). The ultimate objective of CLARIN is to create a European federation of existing digital repositories that include language-based data, to provide uniform access to the data, wherever it is, and to provide existing language and speech technology tools as web services to retrieve, manipulate, enhance, explore and exploit the data. The primary target audience is researchers in the humanities and social sciences and the aim is to cover all languages relevant for the user community.

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 German resources, tools and their integration through web services. Besides these localization efforts, D-SPIN has a special focus in addressing potential users of the infrastructure with the preparation of training material and teaching activities.

This report summarizes the requirements and expectations of the user community with regard to the D-SPIN infrastructure. The data are still preliminary, though, as some of the methods for requirements analysis necessarily rely on existing prototype implementations.

1 Introduction and Context

In the past 20 years textual and linguistic resources have become increasingly useful for all areas of empirically based humanities. The goal of the CLARIN and D-Spin infrastructure is to provide easy and intuitive access to such resources for a broad scientific audience. The success of a resource infrastructure hinges on the acceptance and satisfaction by the community of prospective users. An early requirements analysis can therefore to a certain extent guide or at least support the development activities.

Therefore, the analysis of the needs and expectations of the prospective users has to precede and accompany the implementation of a prototypical D-Spin resource infrastructure.

In Germany, there are currently several projects which aim at facilitating access and use of language resources and tools with a target audience of researchers in the humanities and social sciences. To mention some relevant examples: TextGrid (http://www.textgrid.de/) finished its first phase with the presentation of a prototypical user interface for the interaction with a critical mass of typical language resources (dictionaries, corpora etc.). They organized some user workshops and collected user feedback on these workshops. eAqua (http://www.eaqua.net/) is a more direct co-operation between system developers and humanities researchers, however on the more narrow fields of ancient sources and text mining. The German section of DARIAH operates on a more abstract and strategic level.

The enterprise of constructing a resource infrastructure for researchers has to deal with several stakeholders, among them:

- system and infrastructure developers,
- resource and tool providers and
- researchers who want to use resources and tools as a means to investigate their research questions.

Each of these groups has its own interests in this enterprise, and the interests might be conflicting. It is one of the goals of WP3 to elicit the interests and expectations of the researchers as prospective users of the infrastructure. Even this is not an easy task:

- Humanities and social science researchers constitute a very heterogeneous group in itself with different degrees of affinity to language resources and the use of technological help. While it is obvious why linguists need large amounts of language data, this may be less obvious for historians etc.
- Communities and sub-communities are not easy to get hold of. The best channels are the professional societies of these communities and discussion forums.
- The expectations of these groups might be led by what is already available (“concordances help me a lot with my work”) or by some visions of a future work place and its impact on a concrete research work (“It could help me a lot with my research on ... if I could ...”)
- The community might be hesitant to answer questions which are brought to them by people they do not know and with only a faint idea of what value they might get back for the time and effort they invest to answer these questions.
- It is unfortunately not in the scope of the preparatory phase of D-SPIN to provide a prototype implementation on the basis of which user feedback could be elicited. Nevertheless, there might be a chance that the individual components, e.g. web services and workflows, can be tested with users. This is within the work plan of WP3.

All these difficulties must be overcome with a mix of methods with which to elicit the information which is needed for a realistic and informative requirements analysis.

2 Related Work

An impressive recent requirements analysis is provided by the RePAH (Research Portals for the Arts and Humanities) project (cf. Brown et al. 2006). In their survey of user needs concerning information portals in the Arts and Humanities, the authors applied a mixture of methods, among others a focus groups study, an online-questionnaire and a deep log-file analysis of selected services. This study focuses on the use of information portals. The D-SPIN infrastructure can also be viewed, at least from a user perspective, as an information portal, even if it will be more than that. So this study is an interesting source of inspiration for our work.

In the Netherlands, user requirements analysis has been conducted under the auspices of the SURF initiative (cf. Kircz 2004). This work is mainly based on explorative expert interviews. The author seeks to elicit the "methodological commons" of the humanities disciplines as a common ground for the development of a resource infrastructure.

An "e-science scoping study", edited by Paul Rayson (cf. Rayson 2006), is also based on expert opinions. In this case, the experts took part in some seminars and the study reflects the outcome of these seminars. The experts are all from the field of linguistics, in other words: the target group is narrower than that of the aforementioned studies.

The approach to bring experts together is also applied by the BAMBOO project, which has recently been launched in the US: *"Bamboo is a multi-institutional, interdisciplinary, and inter-organizational effort that brings together researchers in arts and humanities, computer scientists, information scientists, librarians, and campus information technologists to tackle the question: How can we advance arts and humanities research through the development of shared technology services?"* (cf. http://projectbamboo.uchicago.edu/). The project has already conducted some expert workshops. As an outcome of these efforts, a wiki has been set up through which the requirements for the implementation phase of this projects are collected as a common effort of the community. All documents and the wiki are open to the public.

In Germany, a requirements analysis has been started as early as 1986. Researchers of the PROTEXT project in Heidelberg (cf. Report 1986), including one of the authors of this report, conducted some expert interviews with colleagues from the humanities faculties in order to analyze their needs for decentral and dedicated text processing services and tools. The materials are unfortunately no longer available. Lemnitzer and Zinsmeister, in their introduction to corpus linguistics, provide the transcripts of their interviews with some experts in the field. Even if the scope is very narrow, the experts' answers provide some insight into the work of empirical linguists and their needs and wishes with regard to access software and user interfaces (cf. Lemnitzer & Zinsmeister, chapter 7).

A more recent German effort to elicit the requirements of users from the humanities originates in the TextGrid project (http://www.textgrid.de/). The approach taken by this project is twofold. First, one of the partners provided, as a technical report, very detailed user scenarios that guided the development process (cf. TextGrid Szenarien 2006). Second, the project has implemented a prototype of their architecture and offered labs for users to work with and test this prototype (cf. the project website for further details about the prototype). We are however not aware of any report on the user feedback to this prototype.

Aschenbrenner, Blanke and others (cf. Blanke et a. 2008) discuss trends in the eHumanities in Great Britain and Germany. As an empirical base they analyze quite a few eHumanities projects, most of them initiated in Great Britain. Their conclusions are twofold. First, they argue against a technology driven unified solution which tries to capture the assumed needs of all humanities disciplines. Instead, they opt for "local" solutions. Second, they claim that a technological change in the working context of researchers must interact with a change in the research culture ("Forschungskultur") in these disciplines as one of the success factors of the technological change. In their plea for "local" solutions, the authors are in sharp contrast to e.g. Kircz who tries to define a "common methodological" ground of the humanities disciplines (see above).

To summarize, the studies we have referred to show a mix of methods to elicit the requirements of the community in the context of eHumanities, with direct involvement of experts and usage scenarios being the most frequently used method. The findings are still too heterogeneous to allow for easy conclusions and recommendations concerning the technological support for this very broad community.

3 Methodological Issues

“Requirements analysis in systems engineering and software engineering, encompasses those tasks that go into determining the needs or conditions to meet for a new or altered product, taking account of the possibly conflicting requirements of the various stakeholders, such as beneficiaries or users.” (Wikipedia, article: Requirements analysis).

The most common methods of requirements analysis are:

1. eliciting requirements through e.g. interviews and questionnaires,
2. recording requirements, e.g. by collecting data through log-files or drafting usage scenarios and
3. analyzing requirements, e.g. by classifying them, detecting and resolving conflicts between them.

Given the complexity of the task and a certain kind of contextual underspecification, i.e. there being no demonstrator available at this early stage, we have to employ a mixture of methods for the elicitation and recording of user requirements, among them:

- Usage and user feedback analyses of existing textual and linguistic on-line resources. We will describe our efforts in section 3.1. This is a way of recording requirements. The **advantage** of this method is that we are able to glean very realistic data from this kind of feedback. Users send feedback because they feel the need and, in the majority of cases, a dissatisfaction with the service(s) offered. The major **disadvantage** is that the feedback addresses existing services, i.e. state of the art services and obviously not the services which are envisaged as a result of the D-SPIN project. Secondly, some variables concerning the user groups are not controllable. In most cases we simply know nothing about the users who sent the messages.
- Online questionnaires published on scientific mailing lists. The **advantage** of this method is that we address a more or less well-defined community. Furthermore, the answers and their analysis are controlled by the questions we include in the questionnaire. The **disadvantages** are that a) it is not easy to motivate these communities to respond to questionnaires and b) we most probably reach those members of the community who have a high affinity to technical innovation and the patience to answer a questionnaire. These people might not be representative for our user group. We will present our questionnaire and first results in section 3.2.
- Usage scenarios and supervised case studies providing hands-on experience of the D-SPIN infrastructure. Usage scenarios are to be preferred at an early stage of the project. The **advantage** of a well-drafted usage-scenario is that it starts with a real research need and progresses through steps of solving this research need. The major **disadvantage** is that a usage scenario can be very narrow and it is unclear how representative it is for the needs and requirements of the community at large. The major effort in applying this method lies in a careful analysis of the scenario(s) in terms of the consequences for the infrastructure building. CLARIN WP5 initiated such a gathering of usage scenarios, which led to an interesting pool on which we can draw. However, it turned out that many of the scenarios are either unrealistic (“pie in the sky”) or they are not fleshed-out enough to allow for valuable inferences with respect to the infrastrucuture. We will outline the D-SPIN plans in section 3.3.

- Extensive depth interviews with selected experts of their field. The **advantage** of this method is that the analysis is in our control. Furthermore, we can carefully select interview partners who have some influence on a community or represent a community well. The major **disadvantage** is that the application of this method is very cost-intensive both in terms of preparation and in terms of the interview, which is time consuming for both the interviewer and the interviewee and might even involve some travelling for the interviewer. We will outline our plans for depth interviews in section 3.4.

3.1 Analysis of user feedback to existing services

All D-SPIN partners have long experience in providing textual and linguistically annotated data to a broad audience. The data are typically exposed via web servers (e.g. Wortschatz-Portal, http://wortschatz.uni-leipzig.de/; DWDS-Portal, http://www.dwds.de/) or specialized frameworks (e.g. Cosmas2 client, http://www.ids-mannheim.de/cosmas2/). If only for technical reasons, connection data and requests are recorded and these records (log files) are kept for some period of time at the service providing site. These log files are, for this time span, available for analysis e.g. with data mining techniques. Of course, legal aspects apply: the privacy of the individual users has to be respected, which implies at least anonymization of the log file data, regardless of whether the users use the service anonymously or with a log-in account.

On the basis of server log files and error log files it is relatively easy to detect which types of data are requested most frequently and in some cases one can determine which resources the users expected to find but did *not.* For query based systems (e.g. linguistic corpora) the analysis of the actual queries allows to some extent to infer the underlying research question. There is one major methodological drawback connected to log file-based analysis: no meta data about the user is recorded. Therefore no definite conclusion can be drawn as to whether the requested data is intended for scientific research or for any other reason.

A second type of user feedback is the user-initiated direct communication with the resource provider, typically by e-mail. Here, objections, suggestions, requests, criticism (and sometimes praise) are expressed. Because of the relative rarity of user-initiated direct feedback these data hardly allow for generalizations of user requirements and are not considered here.

During the second year of D-SPIN AP3 will contact the projects' resource providers and try to collect and summarize the user feedback for the existing stand-alone services.

Case study: Usage analysis of the DWDS portal

The DWDS website at http://www.dwds.de/ is – with approximately 5 million page impressions (PI) per month – a widely used digital lexical system that provides lexical information for academic users as well as for a broader audience. Currently, the platform contains four different types of information for a given word (Geyken 2005):

1. The dictionary component contains the full digital version of the “Wörterbuch der deutschen Gegenwartssprache” (WDG, “Dictionary of Present-day German”) published between 1962 and 1977 (Klappenbach et al. (1977) and compiled at the Deutsche Akademie der Wissenschaften; the print version comprises six volumes with over 4,500 pages and contains more than 60,000 headwords (more than 120,000 if compounds are counted separately).
2. The corpus component (currently 800 million tokens in total) comprises newspaper corpora, specialized corpora (e.g. spoken language, language of the former German Democratic Republic GDR), and the DWDS core corpus. The core corpus consists of 100 million tokens (comparable in size to the British National Corpus), equally distributed over time and over the following five text types: journalism (approx. 27% of the corpus), literary texts (26%), scientific literature (22%) and other non-fiction (20%), transcripts of spoken language (5%). The corpus is encoded according the guidelines of the text encoding initiative (TEI P5). It is lemmatized with the TAGH morphology (Geyken & Hanneforth (2006)) and tagged with the

part-of-speech tagger moot (Jurish (2004)) in accordance with the conventions of the Stuttgart-Tübingen-Tagset (STTS, Schiller et al. (1999)). The corpus search engine DDC (Dialing DWDS Concordancer, Sokirko (2003)) supports linguistic queries on several annotation levels (word forms, lemmas, STTS part-of-speech categories), filtering (author, title, text type, time intervals) and sorting options (date, sentence length). Details on the design of the corpora and on the technical background of the corpus tools are given in Geyken (2007).

3. An additional thesaurus component computes synonyms, hyponyms and hyponyms for lexical units on the basis of the aforementioned WDG dictionary data (Geyken & Ludwig (2003)).
4. On the basis of the DWDS core corpus, the collocation component offers several options to compute co-occurrences for a lexical unit according to common statistical measures (mutual information, t-score, and log-likelihood). It does not, however, take into account syntactic relations.

The default view combines the four above-mentioned types of information. This view is requested by human agents almost constantly in 82-84% of all page views. Any other view has to be explicitly selected by the user. The "corpus-view" is requested in around 10%, the "dictionary view" in around 6% of the cases. One can safely assume that the actual frequency of use for corpora and dictionary is higher than the given figures as both are part of the default view as well. The figures make it obvious that there is a considerable demand for specific linguistic data.

Another interesting aspect is the frequency and distribution of the types of search which the search engine supports. We can draw two types of inferences from the data: a) we can assume that they reflect the real user needs and define as a requirement for the design of a search engine to best support the most frequently used types of search; b) we can assume that less frequently used types of search are not chosen because they are too complicated or not well presented and define a more easy and visual support of more complex queries as a requirement and on the other hand invest in the teaching of the proper use of search engine. It has to be investigated further which is the more appropriate inference.

For the DWDS portal the following figures sketch a rough view on the types of searches. The vast majority of queries are single words or multi word units that are requested directly without exploiting any of the search operators the system provides. These cases account for 97.4% of all queries. They are used as lemmas for dictionary look-up and corpus searches. Of the query operators that allow for linguistic specifications of tokens and their positional relations the most prominent ones are part-of-speech specification (0.7%), binary Boolean search operators (AND, OR, 0.7%), suppression of paradigmatic expansion of word forms (0.6%) and the specification of a variable directed distance between two or more tokens (0.6%). Less often the NEAR-operator for undirected token distance is used (0.3%) as is the Boolean NOT-operator (>0.1%).[1]

[1] All operators can appear in free combination, thus the classes do not necessarily sum up to 100%.

3.2 Online Survey

3.2.1. The questionnaire

The questionnaire is an instrument of requirements gathering (see above). It is an instrument which can be widely distributed at reasonable costs. As such it is an instrument which produces, in the ideal case, results from which valid statistics can be gleaned.

The questionnaire which we have designed serves to get a general overview of the usage of linguistic tools and resources within the humanities community. The clear limitation of this questionnaire is that its results describe the status quo of the use of language resources. However, at this early stage of infrastructure building this is a useful starting point.

Another issue is the distribution of such a questionnaire. It is not easy to reach a larger number of researchers. One point to start from is to place the request to fill in such a questionnaire in mailing lists which are established in the community. In many mailing lists, such requests are accepted and lead to an acceptable number of responses.

As a first step we have designed such an online survey (cf. Appendix). The survey was initially announced on the mailing list Gespraechsforschung (conversation analysis). We plan to publish the call for participation for the survey on other humanities-related mailing lists. However, this turned out to be more difficult than expected for moderated lists like H-SOZ-U-KULT, the most important mailing list for German historians. Here, the community appears to be closed and not very welcoming to surveys. E-mails to the list administrators remained unanswered so far.

The following analysis of the first survey results is based on the data of the participants of the mailing list “Gespraechsforschung” (research on (spoken) discourse).

By July 2005, about 700 people were subscribed to the list.[2] The announced survey was open from May 11 to June 12 2009. Of the 67 participants 65 answered our questions in the first week. It took them 17 minutes on average to complete.

In the remainder of this section, the survey results are discussed on the basis of the questions. The survey is divided into three major sections: questions on the participant's actual resource use today (1-6), demands, requirements, wishes as to what functionality should be provided by the resources (7-11) and meta data concerning scientific education, age and skills in electronic text/language processing (12).

3.2.2 First results

Ressource use today

1. Which resources do you use? (“Welche Ressourcen verwenden Sie?”)

   Three comprehensive text archives were suggested to the participants (Google book search, http://www.zeno.org (a huge full text library for German books), Projekt Gutenberg (full text library of texts that are no longer subject to copyright)). Two nominal scales are associated

[2] Unfortunately, more recent figures are not provided by the list managers.

with this question: “I do not use it – I use it sometimes – I use it often” and “it is not valuable – it is valuable – it is very valuable”. In addition, a free form field for additional comments (“I would use it more often if ...”) was provided. There were no suggestions recorded for this field.

As expected, a strong correlation between the two nominal scales showed up. Only Google book search was known to every participant and more than 60% claimed to use it. The other two resources were only known to 37% (Zeno.org) and 66% (Projekt Gutenberg) and used by fewer participants.

2. Which digital encyclopedias do you use? (“Welche digitalen Enzyklopädien verwenden Sie?”)

   This question was closely modeled after question 1 and has identical scales. Wikipedia, Brockhaus, Encyclopaedia Britannica (EB) and Spiegel Wissen (SW) were explicitly mentioned. Again, no additional suggestions were given by the participants.

   Wikipedia was the only encyclopedia known to all participants. None of them claimed not to use it and none judged it a useless resource. Brockhaus was unknown to 12% and only 50% of the participants actually use it. EB and SW both are unknown to a quarter of the participants and actively used by only 30%.

3. Which digital dictionaries/word information systems do you use? (“Welche digitalen Lexika/Wortinformationssysteme verwenden Sie?”)

   Again, the nominal scale is exactly the same as for question 1. The given choice comprised Duden, Deutscher Wortschatz, canoo, DWDS, Grimm's dictionary and Kluge's dictionary. While Duden is virtually known to every participant the other systems are unknown to 30-40% with the exception of Grimm's dictionary (unknown to only 20%). Usage claims correlate with the publicity of the electronic dictionaries. Only very few judgments were made on the utility of these resources. These are therefore ignored in the discussion. No additional suggestions were made by the participants.

4. Which linguistic corpora do you use? (“Welche linguistischen Corpora verwenden Sie?”)

   The nominal scale is the same as for question 1. The given choice comprised the IDS and the DWDS corpora and the TIGER corpus. TIGER was unknown to the majority of participants (64%) and hardly used at all (4%). Of the more general corpora, IDS is more often used (62%) than DWDS (25%) and better known (88% vs. 64%). Other corpora have not been mentioned by the participants.

5. How do you use the resources? (“In welcher Form recherchieren Sie in den Ressourcen?”)

   Here, the nominal scale allowed for the unrestricted selection of “facsimile view”, “full text search” and “linguistic search”. The most prominent use appears to be full text search (73%) whereas both the facsimile view on the data and the search on linguistically annotated layers achieve a rate of 50%.

6. Which kind of linguistic search do you use?) (“Welche Art linguistischer Suche verwenden Sie?”)

   The choice for the participants was between “concordance lines”, “morphological search/lemmas”, “part-of-speech search”, “search for semantic relations”, “named entity search” and “search for syntactic patterns”. Multiple choices were allowed and a field for further properties was provided. Each choice was selected between by 15% and 30% with semantic relations being most and named entities least prominent.

   A considerable amount of additional suggestions was made by the participants most of which can be motivated by the shared research subject of the mailing list members (conversation analysis): phonological search, multi-layer searches (e.g. phonological form and phonetic realization (repeatedly and independently nominated)), paraverbal entities, extraverbal entities, multi-modal data and also “grammatical rules” and etymological information.

Requirements for future systems

7. What functions should text archives/corpora make available? (“Welche Funktionalität sollten Volltextsammlungen/Corpora bieten?”)

   The nominal scale for this question contained the items “full text download”, “possibility to edit the texts” and “notes attachment”. The most important functions are “full text download” (76%) and the possibility to attach notes (50%).

   Many of the additional suggestions are again due to the topic of the mailing list: viewing and downloading of audio/video samples (multiple times); gestures/interaction patterns; copy and paste/export into other applications; personalized searches, re-use of previously created result sets and full text search.

8. Which export formats should text archives/corpora provide? (“Welche Ausgabeformate sollten Volltextsammlungen/Corpora unterstützen?”)

   The explicit nominal scale contained “TXT” (52%), “XML” (37%) and “PDF” (72%). All seven suggestions by the participants exclusively focused on word processor formats (“DOC”, “RTF”).

9. Which means of text editing should text archives/corpora provide? (“Welche editorischen Möglichkeiten sollten Volltextsammlungen/Corpora haben?”)

   The nominal scale contained “orthographic normalization”, “named entity annotation” and “structural annotations”. This question was largely ignored (88%) and will therefore not be discussed in detail. There was one additional suggestion to include one's own categorization.

10. Can you imagine using web interfaces for linguistic tools? (“Können Sie sich die Nutzung von Webschnittstellen für linguistische Tools vorstellen?”)

    The nominal scale comprised “CGI/form data” and “service based (e.g. XML-RPC, SOAP)”. 51% of the participants nominated “CGI/form data” and 22 % “service based”. There were two remarks as to not knowing what the question was actually about and 43% of the participants simply skipped it silently.

11. What amount of text would you like to process using a web interface? (“Welche Textmengen würden Sie mit Hilfe einer Webschnittstelle bearbeiten wollen?”)

    The question was associated with a nominal scale of “less than 10 documents”, “10-100 documents”, “more than 100 documents” and “I don't know”. 32% of the participants answered the question with the majority voting for “10-100”; 24% skipped the question and 43% explicitly marked “I don't know”.

Participants' meta data

12. Meta data and personal profiles.

    Based on the topic of the mailing list one can reasonably assume that all participants have a background in linguistics and more specifically in conversation analysis. This was confirmed by the question regarding the area of expertise. Virtually every participant answered the self assessment question on word processor skills with “good” to “very good” while only 35% had the same impression of their XML editor skills.

Discussion

The participants of the pilot survey form a community with tight bonds into linguistics and a strong demand for empirical linguistic data.

The survey reflects two of the basic motivations for the CLARIN/D-Spin initiative. There are very few huge generic resources (e.g. Google book search, Wikipedia) that are well known across the field and that are used on a daily basis. Medium sized resources (e.g. canoo, DWDS, Projekt

Gutenberg) are used significantly less frequently while the smaller resources tend to be overlooked completely (e.g. TIGER). A common integrating infrastructure will certainly help the community to discover the wealth of resources that are available to them today. Secondly, the technical state of the art of the majority of the participants becomes apparent: there is a high demand of data in formats that do not fit very well into any linguistic tool chain (cf. question 8: word processor formats, PDF). Also, command of word processors is far better than that of XML editors. Obviously, the D-Spin infrastructure implementation has to become as user friendly as possible and hide the technical particularities of the system from the users as far as possible. Training of the user communities and providing technical support for them will be crucial points in the future of D-Spin.

3.3. Usage Scenarios

As mentioned above, the CLARIN project has already conducted a survey of usage scenarios on the European level. The results of this survey can be found at: http://www.clarin.eu/wg-51/usage-scenarios (note that you have to be registered and logged in to view this page). There is a CLARIN deliverable available which analyzes these usage scenarios.

It has been decided on a community meeting in Mannheim, (15 and 16 May 2009) to launch a similar survey addressed to the German humanities community. First steps have been taken. However, some problems are still to be solved before this survey can be launched and announced.

The narrow focus of the usage scenario approach is both the strength and the weakness of it. The strength is that a usage scenario can be fleshed out to the very details and thus lead to very precise specifications for the infrastructure. The risk is that a small set of usage scenarios does not reflect the needs of the user community very well. They have therefore to be accompanied by depth interviews with experts in the field(s).

3.4. Planning of explorative expert interviews

In addition to the online survey we will conduct explorative expert interviews with carefully selected specialists and key persons of different fields within the humanities. The questions will be based on the findings of the online survey and on usage scenarios. We plan to conduct interviews with a small list of experts in the following areas:

- Historical semantics: T. Gloning (Univ. Giessen)
- Philology: M. Niedermeier, (Goethe-Wörterbuch, BBAW), F. Martin (Corpus Vitrearum Medii Aevi, BBAW)
- Educational sciences : A. Storrer (Univ. Dortmund), H. Drumbl (Univ. Bozen)
- Linguistics: M. Stede (Univ. Potsdam), E. Burr (Univ. Leipzig)

The interview questions will cover the following topics:

1. the expert's view of the need of language resources in the community he or she represents, probably in contrast with and comparison to the findings of the online survey,
2. the expert's own research and the role which language resources can play or could play in it,
3. the current scientific practice with respect to language resources in the field and
4. the expert's vision of the research process in the future and what is required with respect to the availability and usability of language resources and tools.

An important part of the interviews will be guided usage scenarios based on prototypical implementations of web platforms. Here, we plan to use an implementation of the D-Spin prototype infrastructure or a mock-up of such a system. We also plan to use other comparable platforms such as the TextGrid lab and the web archive of the German Text Archive (Deutsches Textarchiv-Prototype, http://www.deutsches-textarchiv.de/, available by September 2009). The latter platform will focus on cumulative work with text resources and as such will be an interesting demonstrator for some research fields.

4 Outlook

We will continue our efforts in collecting requirements for the emerging LRT infrastructure. In particular, we will a) present our questionnaire to additional communities by launching it in more discussion forums; b) prepare and conduct expert interviews, c) collect user feedback to the workplace of the "Deutsches Textarchiv", which is one good example for cumulative work with corpora, and d) prepare an expert workshop which will probably take place in spring 2010. Therefore, the milestones for year 2 of the project will be:

- October 2009: an updated analysis of the questionnaire
- December 2009: a report which summarizes the findings of the expert interviews
- March 2010: Expert workshops on visions about the eScience research environment for humanities researchers
- December 2010: a report about user feedback to the DTA workplace

References

Andelfinger, U. (1997): *Diskursive Anforderungsanalyse. Ein Beitrag zum Reduktionsproblem bei Systementwicklungen in der Informatik*. Frankfurt et al.: Peter Lang.

Aschenbrenner, A. & Blanke, T. & Dunn, S. & Kerzel, M. & Rapp, A. & Zielinski, A. (2007): *Von e-Science zu e-Humanities – Digital vernetzte Wissenschaft als neuer Arbeits- und Kreativbereich für Kunst und Kultur*. In: *Bibliothek. Forschung und Praxis*, 1, S.11-21.

Blanke, T. & Aschenbrenner, A. & Küster, M. & Ludwig, C. (2008): *No Claims for Universal Solutions – Possible Lessons from Current e-Humanities Practices in Germany and the UK*. In: *e-Humanities – An Emerging Discipline.* Workshop at the 4th IEEE International Conference on e-Science. December 2008.

Brown, S. & Ross, R. & David G. & Greengrass, M. & Bryson, J. (2006): *RePAH: A User Requirements Analysis for Portals in the Arts and Humanities*. Final Report. (http://repah.dmu.ac.uk/report/index.html)

Geyken, A.; Ludwig, R. (2003): *Halbautomatische Extraktion einer Hyperonymiehierarchie aus dem Wörterbuch der deutschen Gegenwartssprache* [on-line]. TaCoS 2003. (http://kollokationen.bbaw.de/doc/ExtrHyp.pdf).

Geyken, A. (2005): *Das Wortinformationssystem des Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts (DWDS)*. In: *BBAW Circular 32*. Berlin: BBAW.

Geyken, A.; Hanneforth, T. (2006): *TAGH – A Complete Morphology for German based on Weighted Finite State Automata*. *Proceedings of FSMNLP 2005*. 55-66.

Geyken, A. (2007): *A reference corpus for the German language of the 20th century.* In Fellbaum C. (ed.). *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London: Continuum Press.

Gogner, A. & Littig, B. & Menz, W. (2005) (eds.): *Das Experteninterview. Theorie, Methode, Anwendung*. 2. Auflage, Wiesbaden:Verlag für Sozialwissenschaften.

Jurish, B. (2003): *Part-of-Speech Tagging with Finite State Morphology*. Poster presented at the conference Collocations and Idioms: Linguistic, Computational, and Psycholinguistic Perspectives, Berlin, 18.-20. September, 2003.

Kircz, J. (2004): *E-based Humanities and E-humanities on a SURF platform*. A SURF-DARE Technical Report, June 2004 (http://dare.uva.nl/record/119979).

Klappenbach, R.; Steinitz, W. (eds.) (1964-1977): *Wörterbuch der deutschen Gegenwartssprache (WDG)*. Berlin: Akademie-Verlag.

Lemnitzer, L. & Zinsmeister, H. (2007): *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

Maciaszek, L. (2007): *Requirements Analysis & Systems Design*. Addison Wesley.

Rayson, P. (2006): *AHRC e-science Scoping Study Final Report: Findings of the Expert Seminar for Linguistics.* (ahds.ac.uk/e-science/documents/Rayson-report.pdf).

Reports from Colloquia at Tübingen (1986). In: *Literary and Linguistic Computing* 1, 3, p. 176-178.

Rupp, C. (2007): *Requirements-Engineering und -Management: Professionelle, iterative Anforderungsanalyse für die Praxis*. München:Hanser.

Sokirko, A. (2003): DDC. In: *Computational linguistics and intellectual technologies.* Protvino, Russia.

TextGrid Szenarien. Technischer Report des TextGrid Projekts. December 2006 (www.textgrid.de).

Online Questionnaire

The online questionnaire is appended to this report.

Vorwort

Umfrage zur Nutzung von sprachlichen Ressourcen (Textkorpora, Wörterbücher u.Ä.)

Wir möchten Sie einladen, an einer Umfrage zur Nutzung von sprachlichen Ressourcen (Textkorpora, Wörterbücher u.Ä.) teilzunehmen.
Als ein Ziel dieser Erhebung möchten wir ermitteln, welchen Ansprüchen eine Infrastruktur, in der sprachliche Ressourcen der Forschung zur Verfügung gestellt werden, genügen sollte.

Wer sind wir?

Die Umfrage geht aus vom Projekt D-SPIN (www.sfs.uni-tuebingen.de/dspin). D-SPIN ist der deutsche Beitrag zum europäischen Projekt CLARIN (Common Language Resources and Tool Infrastructure, www.clarin.eu) und hat zum Ziel, auf nationaler Ebene eine Sprachressourcen-Infrastruktur für den geistes- und sozialwissenschaftlichen Arbeitsplatz der zukunft zu schaffen.
Mit der Teilnahme an dieser Umfrage helfen Sie uns, die Anforderungen, die diese Infrastruktur erfüllen können sollte, um für Ihre Forschungen nützlich zu sein, besser zu verstehen. Dies betrifft sowohl die Inhalte, die verfügbar gemacht werden sollten, als auch die Art und Weise des Zugriffs und der Nutzung dieser Ressorcen.

Bitte nehmen Sie sich ca. 10 Minuten Zeit, um die 12 Fragen zu beantworten. Die Ergebnisse der Umfrage werden wir auf der Webseite des D-SPIN-Projekts veröffentlichen.

Frage 1

Welche Ressourcen verwenden Sie?

Maximal 2 Antworten pro Zeile

	Verwendung und Nutzen bei der Arbeit							würde ich (öfter) verwenden, wenn ...
	verwende ich nie	verwende ich gelegentlich	verwende ich häufig	ist nicht nützlich	ist nützlich	ist sehr nützlich	ist mir unbekannt	
google books	□	□	□	□	□	□	☑	
zeno.org bzw. Digitale Bibliothek	□	□	□	□	□	□	☑	
Projekt Gutenberg	□	□	□	□	□	□	☑	
weitere	□	□	□	□	□	□	□	
weitere	□	□	□	□	□	□	□	
weitere	□	□	□	□	□	□	□	

Frage 2

Welche digitalen Enzyklopädien verwenden Sie??

Maximal 2 Antworten pro Zeile

	Verwendung und Nutzen bei der Arbeit							würde ich (öfter) verwenden, wenn ...
	verwende ich nie	verwende ich gelegentlich	verwende ich häufig	ist nicht nützlich	ist nützlich	ist sehr nützlich	ist mir unbekannt	
Wikipedia	□	□	□	□	□	□	☑	
Brockhaus	□	□	□	□	□	□	☑	
Britannica	□	□	□	□	□	□	☑	
Spiegel Wissen	□	□	□	□	□	□	☑	
weitere	□	□	□	□	□	□	□	
weitere	□	□	□	□	□	□	□	
weitere	□	□	□	□	□	□	□	

Frage 3

Welche digitalen Lexika/Wortinformationssysteme verwenden Sie?

Maximal 2 Antworten pro Zeile

	Verwendung und Nutzen bei der Arbeit							würde ich (öfter) verwenden, wenn ...
	verwende ich nie	verwende ich gelegentlich	verwende ich häufig	ist nicht nützlich	ist nützlich	ist sehr nützlich	ist mir unbekannt	
Duden	□	□	□	□	□	□	☑	
Deutscher Wortschatz/Leipzig	□	□	□	□	□	□	☑	
canoo	□	□	□	□	□	□	☑	
DWDS	□	□	□	□	□	□	☑	
Grimmsches Wörterbuch	□	□	□	□	□	□	☑	
Kluge	□	□	□	□	□	□	☑	
weitere	□	□	□	□	□	□	□	
weitere	□	□	□	□	□	□	□	
weitere	□	□	□	□	□	□	□	

Frage 4

Welche linguistischen Corpora verwenden Sie?

Maximal 2 Antworten pro Zeile

	Verwendung und Nutzen bei der Arbeit							würde ich (öfter) verwenden, wenn ...
	verwende ich nie	verwende ich gelegentlich	verwende ich häufig	ist nicht nützlich	ist nützlich	ist sehr nützlich	ist mir unbekannt	
IDS-Mannheim	□	□	□	□	□	□	☑	
DWDS-Korpus	□	□	□	□	□	□	☑	
TIGER	□	□	□	□	□	□	☑	
weitere	□	□	□	□	□	□	□	
weitere	□	□	□	□	□	□	□	
weitere	□	□	□	□	□	□	□	

Frage 5

In welcher Form recherchieren Sie in den Ressourcen?

Mehrfachantworten möglich

□ Einfache Bildanzeige digitalisierter Bücher/Zeitschriften

□ Volltextsuche

□ Linguistische Suche

Frage 6

Welche Art Linguistischer Suche verwenden Sie?

Mehrfachantworten möglich

□ Konkordanzen

□ Lemmasuche, Morphologische Suche

□ Suche nach Wortart

□ Suche nach semantischen Relationen (z.B. Synonyme)

□ Suche nach Eigennamen

□ Suche nach syntaktischen Mustern

□ Andere:

Welche anderen Arten der Suche verwenden Sie (offene Frage)?

Frage 7

Welche Funktionalität sollten Volltextsammlungen/Corpora bieten?

Mehrfachantworten möglich

□ Download des Volltexts

□ Möglichkeit, eigene editorische Änderungen an den Texten durchzuführen

□ Notizzettel-Funktion

□ Andere:

Welche anderen Funktionalitäten würden Sie sich wünschen (offene Frage)?

Frage 8

Welche Ausgabeformate sollten Volltextsammlungen/Corpora unterstützen?

Mehrfachantworten möglich

□ TXT

□ XML

□ PDF

□ Andere Formate:

Weitere Bemerkungen zu dieser Frage:

□

Frage 9

Welche editorischen Möglichkeiten sollten Volltextsammlungen/Corpora haben?

Mehrfachantworten möglich

□ Schreibweisennormalisierung

□ Annotation von Eigennamen

□ strukturelle Annotation (z.B. Reimstruktur, rhetorische Figuren etc.)

Frage 10

Können Sie sich die Nutzung von Webschnittstellen für linguistische Tools vorstellen?

Mehrfachantworten möglich

□ Als Normale Webschnittstelle/Formular

□ Als Servicebasierte Schnittstelle, z.B. XML-RPC oder SOAP

□ Ich würde keine Webschnittstellen nutzen

Frage 11

Welche Textmengen würden Sie mit Hilfe einer Webschnittstelle bearbeiten wollen?

○ weniger als 10 Dokumente
○ 10-100 Dokumente
○ mehr als 100 Dokumente

○ Weiss nicht/Keine Angabe

Frage 12

Bitte vervollständigen Sie die folgenden Angaben:

Ich studiere --- ▼

Ich schreibe gerade eine Qualifikationsarbeit im Bereich

Ich forsche im Bereich

Alter --- ▼

Einschätzung Ihrer Computerkenntnisse

Werkzeuge Textverarbeitung --- ▼

Werkzeuge XML-Editoren --- ▼

Werkzeuge linguistischen Annotation --- ▼

Danke!

Umfrage zur Nutzung von sprachlichen Ressourcen (Textkorpora, Wörterbücher u.Ä.)

Vielen Dank für Ihre Mithilfe