Erschienen in: Proceedings des Workshops TMS – Text Mining Services, Leipzig. Leipzig: 2009. S. 1-15.

Language Resources, Taxonomies and Metadata

Lothar Lemnitzer[1,2] and Erhard Hinrichs[1] and Andreas Witt[3]

[1] Seminar für Sprachwissenschaft, Universität Tübingen, Germany
[2] Berlin-Brandenburgische Akademie der Wissenschaften, Berlin, Germany
[3] Institut für Deutsche Sprache, Mannheim, Germany

Abstract. In this paper we present an approach to faceted search in large language resource repositories. This kind of search which enables users to browse through the repository by choosing their personal sequence of facets heavily relies on the availability of descriptive metadata for the objects in the repository. This approach therefore informs the collection of a minimal set of metatdata for language resources. The work described in this paper has been funded by the EC within the ESFRI infrastructure project CLARIN[4].

1 Introduction

The ultimate objective of the ESFRI infrastructure project 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 be stored, 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. The objective of the current CLARIN Preparatory Phase Project (2008-2010) is to lay the technical, linguistic and organisational foundations, to provide and validate specifications for all aspects of the infrastructure (including standards, usage, IPR) and to secure sustainable support from the funding bodies in the (now 23) participating countries for the subsequent construction and exploitation phases beyond 2010.

One of the major tasks of the CLARIN project is to produce a broad and detailed survey of language resources and tools. The rationale for this task is threefold:

1. it serves as the empirical basis for the prototypical integration of resources and tools into an emerging web service infrastructure;
2. it serves as the empirical basis for the development of a Basic Language Resource tool Kit (BLARK), that primarily serves the research needs of humanities and social science scholars;

[4] Grant agreement no. 212230; we are grateful to the EC for the generous funding without which the reported work would not have been possible. We also would like to thank Peter Wittenburg and two anonymous reviewers for their helpful remarks on earlier versions of this paper.

3. it serves as the empirical basis for assessing to what extent existing language resources conform to existing standards for language resources and metadata

Besides these three project-internal reasons the data collected in this survey will of course constitute a valuable, freely accessible resource for any potential users of LRs and tools, especially for humanities and social science researchers.

The purpose of this paper is to specify the requirements for user-friendly ways of accessing information about sets of language resources that are of interest to humanities and social science scholars. It surveys existing web-based front-ends, such as registries and catalogues, and suggests ways of enhancing the existing functionalities of such front-ends. Such functionalities are intimately tied to the types of metadata that are available for individual language resources. So as to not overburden resource providers with unnecessarily complex and detailed questionnaires for metadata collection, a minimum set of metadata needs to be defined. The specification of this set will be partly informed by the predefined categories and hierarchies and will otherwise conform to metadata standards such as TEI header (Text Encoding Initiative, cf. `www.tei-c.org`), Dublin Core cf. `http://dublincore.org/`) and IMDI (ISLE Metadata Initiative, cf. `http://www.mpi.nl/IMDI/`).

This paper is related to the efforts on the development of the CLARIN registry infrastructure. It therefore addresses the efforts towards an interoperable language resource and metadata federation from the point of view of the language resource infrastructure building. The purpose of a metadata federation infrastructure is to provide information about and access to available language resources and services in a systematic fashion. The development of such a federation, thus, represents an important prerequisite for the interoperable infrastructure of language tools and resources that CLARIN aims to develop.

It is important for the target users to be able to navigate, i.e. browse and query, a complex repository of tools and resources. This presupposes that the tools and resources are categorized in a uniform manner and that the categories provide the basis of conceptual hierarchies that best reflect the information and research needs of the potential users of the CLARIN infrastructure. This naturally leads to the idea of providing different views or facets of the available tools and resources along different dimensions. Such facets are likely to include the different languages, modalities of language (spoken, written, multimodal), intended user groups (e.g. linguists, historians, social scientists) and contexts of use (e.g. information retrieval, machine translation, and speech recognition).

Our work will be informed by existing efforts and initiatives for cataloguing and categorizing language tools and resources. Therefore, in section 2 we will present and review prominent examples of existing registries (DFKI, ACL) and of existing catalogues (ELRA) as well as the IMDI metadata browser developed by the Max Planck Institute for Psycholinguistics in Nijmegen. In section 3, we will present a detailed proposal of a set of hierarchically structured views of tools and resources.

At the moment, several initiatives are actively working on registry structures. The important work in this field is carried out by the ISO TC37 and by the Max

Planck Institute for Psycholinguistics in Nijmegen. Some relevant background information on the state of the ISO approach can be found in:

- ISO TC37/SC4: Infrastructure Note on Registry Databases[5];
- Data Category Registry DCR2 Requirements[6].

The *Infrastructure Note on Registry Databases* introduces a separation of registries into registries of the resources and relation registries. Given this separation, the views of tools and resources would be found in one or more relation registries, whereas the registry of resources contains all the metadata directly associated to the resources.

2 What web-based front-ends are available?

In the field of computational linguistics (CL) a wide variety of tools and frameworks for processing natural language have been and will be developed. Since an influential branch in CL also deals with the development and the application of statistical methods, there is also a need for tools to access and process large quantities of data. Moreover, language resources, especially corpora and lexica, play an important role in linguistic research. In response to a rising need to get an overview of existing methods, tools, resources and frameworks, some institutions and organisations gathered information on tools and resources in registries and made it available on the web. These registries should be considered as building blocks or reference models for the CLARIN registry. The main purpose of the existing registries is to support humans in finding what they are looking for. This will also be an important use case of the search and browsing facilities on federated metadata provided by CLARIN. The CLARIN infrastructure will also support the searching for resources by non-human agents, e.g. automatic processes. One feature that all the registries we investigated share is that they either do not provide multiple views on the same resources or provide them only in a very limited way. This means, that the tools and resources catalogued in the existing registries are not browsable according to several different perspectives arising from different information needs. The infrastructure we are going to establish will not only allow the presentation of the resources along multiple views. It should rather be possible for users to browse the repository by drilling into several category trees to arrive at a level of specifity which they consider to be sufficient for their information needs.

2.1 DFKI software registry

The Natural Language Software Registry (NLSR, cf. `http://registry.dfki.de/`) was developed and set up by the German Research Institute on Artificial Intelligence (DFKI), which is also a CLARIN partner. The DFKI registry

[5] cf. [4].
[6] cf. [3].

presents information on natural language processing software. It lists academic as well as commercial software. The NLSR can be seen as a taxonomy, hence its structure is a tree. At the leaves, the products are listed. Each product is associated with a set of information, amongst them the languages for which the software can be used, the terms on which it can be acquired, its price and the name of a contact person. In addition to the NLP software, NLSR also lists some other Natural Language Resources, e.g. corpora, but only to a limited extent. The NSLR does not consider it to be in their focus to present an exhaustive overview of resources. The users who are interested in these resources are referred to other institutions, especially the ELRA (see section 2.3). On the left-hand side of the start page and on all the other pages of this website a navigation panel is shown (see fig 1). It allows the user to navigate to several areas. The user can submit new resources (item 4 in the navigation panel) or search for resources (item 3). The taxonomical structure is accessible through item 2.

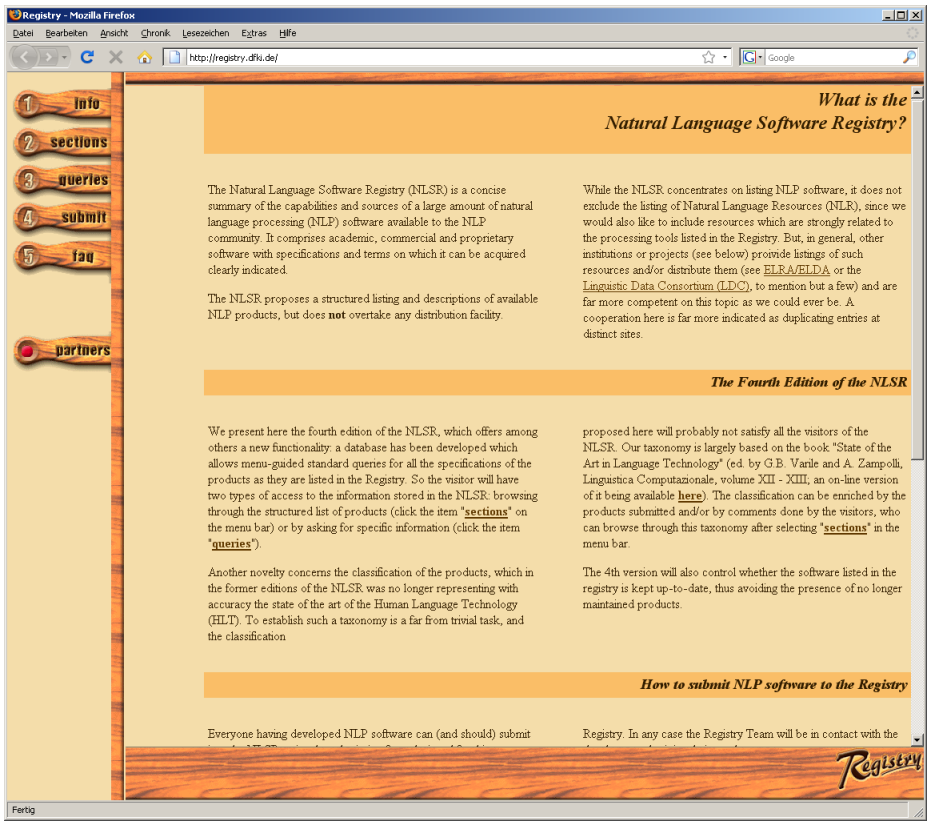

Fig. 1. The DFKI Software Registry

The tools are grouped into eight categories, e.g. tools for (manually) annotating resources, tools for processing of spoken or written language data, evaluation tools etc. The groups *Annotation tools* and *Multimedia* do not contain any sub-groups. *Evaluation tools* is subdivided in three categories, i.e. *Evaluation of Machine Translation, Evaluation of Parsers* and *Evaluation of Speech Synthesis*. The category *Language Resources* does not contain corpora and resources, but tools to process resources. It is subdivided into *Grammar Resources, Lexica, Multimodal Corpora, Spoken Language Corpora, Terminology Tools* and *Written Language Corpora*. Tools dealing with multimodal resources are categorised into the following four groups: a) Facial Movement and Speech Recognition, b) Facial Movement and Speech Synthesis, c) Speech and Gesture and d) Text and Images. Yet another category deals with software which is used by computational linguists to develop and implement NLP-systems. Since this group of tools is not in the scope of CLARIN, we will not mention its sub-categories here. The largest groups of tools are listed under to the headings 'Spoken Language' and 'Written Language'. The group of spoken language tools is subdivided as follows:

```
Signal Editing
Signal Processing
Sound Change Simulation
Speaker Recognition
Speech Analysis
Speech Editing
Speech Processing Applications
Speech Production
Speech Recognition Applications
Speech Synthesis Applications
Spoken Dialog Environments
Spoken Language Generation
Spoken Language Translation
Spoken Language Understanding Applications
Text-to-Speech Synthesis
Voice Analysis
Voice Control
Voice Dialing
Voice Processing
```

The tools for processing written language are sub-grouped as follows (to save space we present only a part of the hierarchy, for a full list cf. `http://registry.dfki.de/`):

```
Alignment Tools
Comparative Linguistics
Controlled Language Applications
Corpus Analysis
Deep Generation Applications
Deep Syntactic Analysis
```

```
Document Image Analysis
Dynamic Hyperlinking
Grammar Checking Applications
...
```

When clicking on a category all tools belonging to this group are listed. The number of the tools in a category is displayed in parentheses immediately after the name of the category. In general, tools are categorised as belonging to several (sub)categories.

2.2 ACL Web registry

The ACL web registry is a wiki-system provided and hosted by the Association for Computational Linguistics (CL), one of the most important societies of the world-wide CL community, cf. `http://aclweb.org/aclwiki/index.php?title=Resources`. In contrast to the DFKI registry, which has been established and is being maintained in a centralized way by a project devoted to this task, the ACL web registry is a community effort. It is the language resource providers and software developers themselves who are responsible for making their resources known to the public through this registry. A screen shot of the registry's top level taken in February 2009 shows, for English, the resource types *corpora, dictionaries, generation grammars, geographical words, knowledge collections and data sets, lexicons, subject specific resources, tools and software* and *uncategorized resources*. Fig. 2 is a screen shot of the ACL registry cover page.

Beside these categories there is also a second taxonomy *List of resources by language*[7]. Furthermore, the categorisation differs between languages. For example, the top level categories for German are *Corpora, Evaluation datasets, Grammars, Lexicons, Resource Access, and Timeline Analysis* and the ones for Greek are *Machine translation systems, Corpora, Named entity recognition, Natural language generation, part of speech tagging* and *Bibliography*. The main reason for these differences lies in the bottom-up, decentral organisation of wikis. Users are free to define their categories as they like. Some of the ACL categories are further subdivided. To give an example of structure at a subordinate level, the categories of the section *Tools and Software* for English is presented here:

```
Cognate identification software
Dialectometrics software
Educational software
Information retrieval software
Knowledge representation software
Lexicon extraction software
Machine translation software
Morphology and part of speech tagging
Multilingual software
```

[7] cf. `http://aclweb.org/aclwiki/index.php?title=List_of_resources_by_language`.

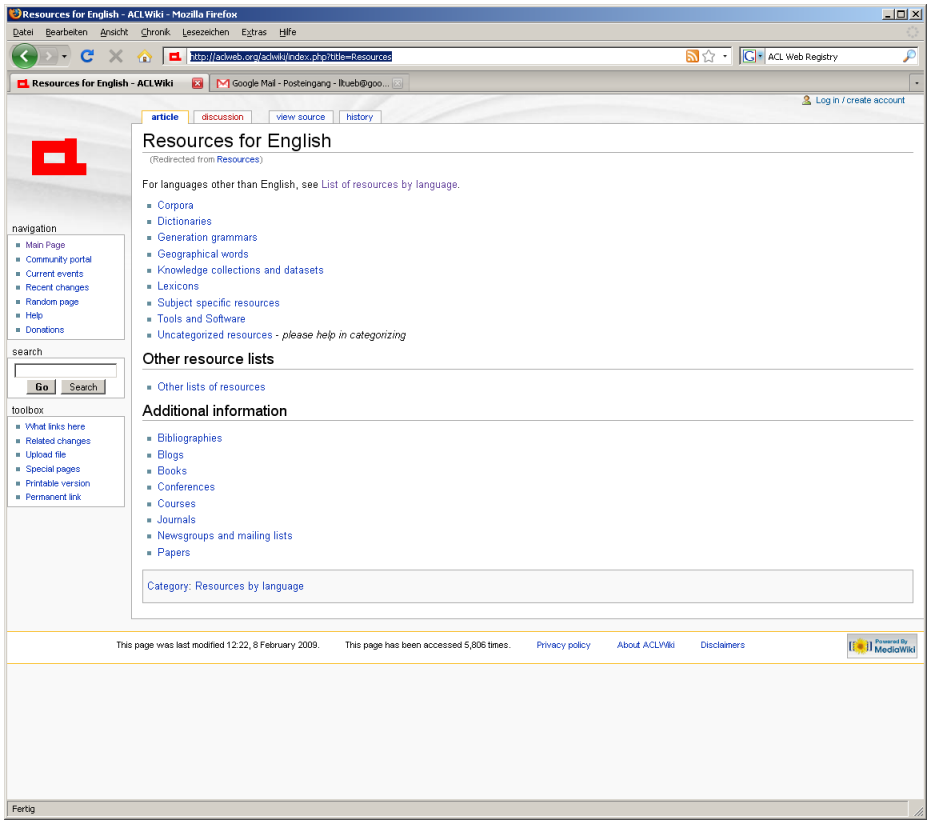

Fig. 2. The ACLWeb Registry

```
Named entity recognizers
Natural language generation software
Natural language interfaces
Phonology software
Parsers
Semantics software
Speech software
Syntax and grammar
```

Of course, also these sub-structures are not applied consistently, but differ between languages.

2.3 The ELRA catalogue

The European Language Resources Association (ELRA) aims at making available language resources of various kinds for a wide range of languages. To find

language resources for purchase, a catalogue is provided by ELRA which is accessible via WWW front-end (cf. 3)[8]. The structure of the ELRA catalogue is kept quite simple. At a top level the resources are grouped into four categories, i.e. spoken, written, terminological and multimodal language resources. Terminological and multimodal language resources are not further subdivided. Written resources are grouped into the three categories: *Corpora, Monolingual lexicons, and Multilingual lexicons*. ELRA's catalogue for spoken resources is further subdivided into *Telephone recordings, Desktop/Microphone recordings, Broadcast Resources* and *Speech Related Resources*. In the browsing mode, the ELRA catalogue does not offer a "browse by language" feature. The searching by language can be evoked by submitting a general query.

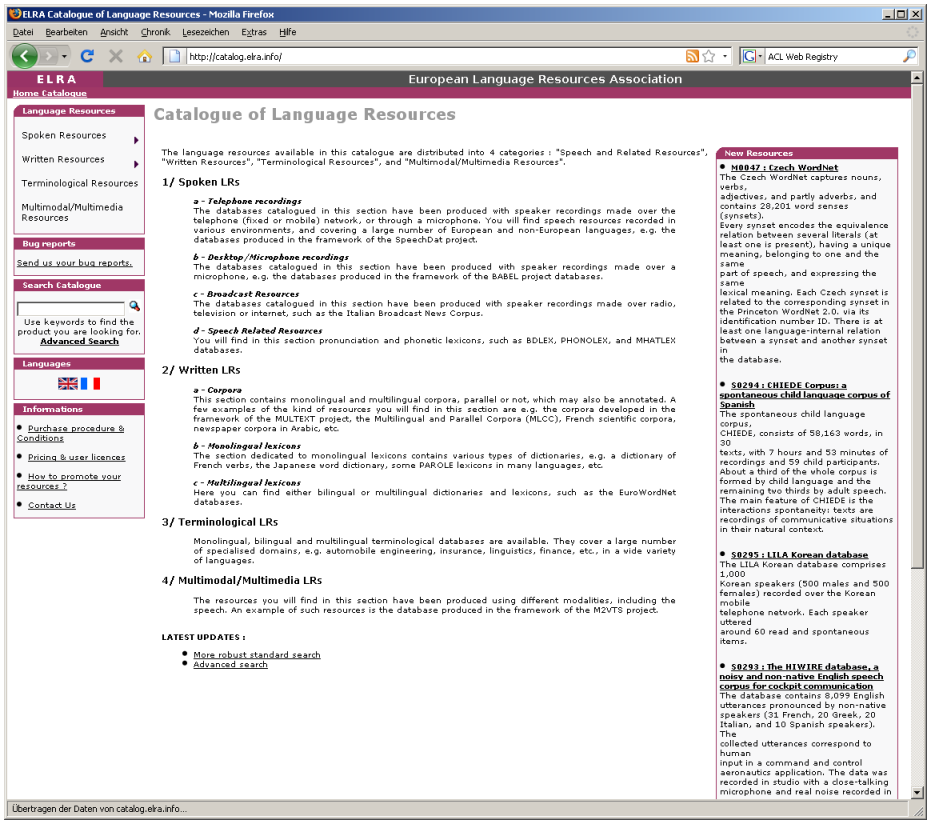

Fig. 3. Front page of the ELRA catalogue

[8] The ELRA catalogue is now extended by a so-called *Universal Catalogue*, cf. *http://universal.elra.info/*. The latter is, however, accessible by a search interface only and is therefore not relevant in our context.

2.4 The MPI tools for IMDI-Data

Language resources using the IMDI metadata scheme can be accessed through the IMDI browsers developed by the MPI Nijmegen. A central functionality which distinguishes the IMDI tools from the web interface provided by the above mentioned resource collections is the support for providing hierarchical access to the resources. This can be done by organizing the resources in the form of tree structures, consisting of nodes that group files together. Resources can be freely grouped by the provider of the resource collection. This grouping could be based on, e.g., the geographical region, the discourse genre, the sex or age of subjects etc. The IMDI Tree Builder is a tool that supports the creation of these hierarchies. It allows for creating trees based on IMDI metadata descriptors. To display these structures the IMDI Tree Browser is used. Figure 4 shows a web access to IMDI data. The tree structure is displayed at the left-hand side of the editor.

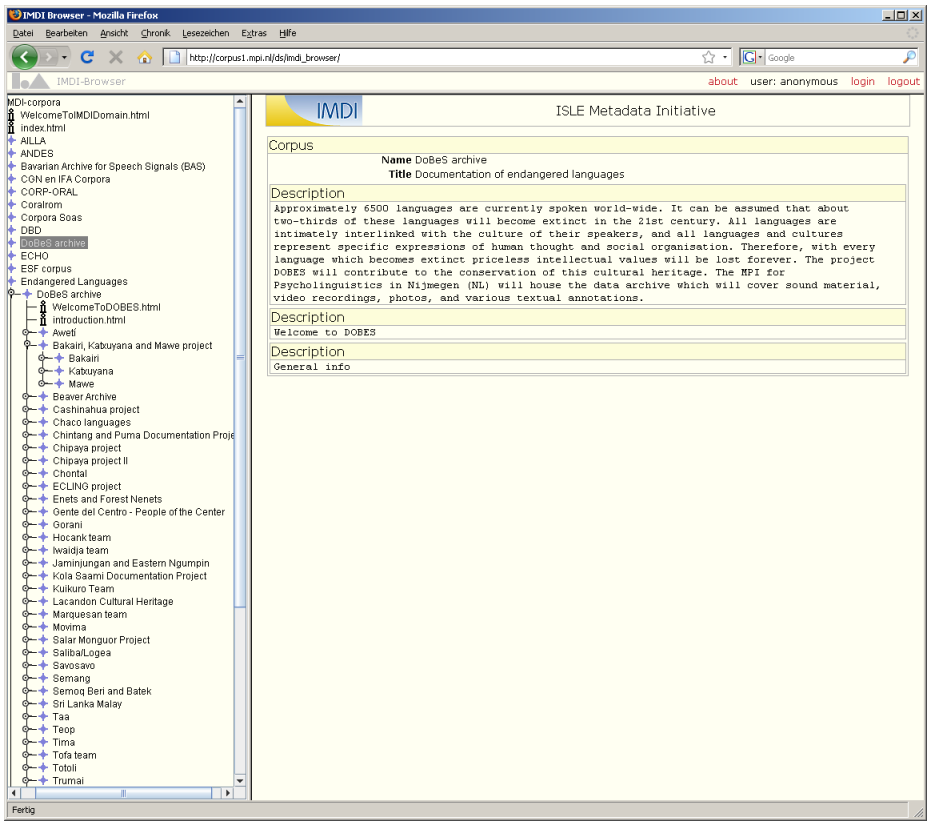

Fig. 4. The IMDI Browser

The next section lists several views on language resources and tools. Ideally, all of these views should serve as potential access points to the resource collections. The views presented in the following section have been drafted based on the knowledge obtained through the analysis of the registries described above, by their own knowledge of the field and by their own experience as linguistic researchers. The IMDI browser could serve as a tool for retrieving tools and resources according to different predefined modes[9]. Predefined hierarchies of the language resources and tools must be established by the CLARIN project.

3 Views and facets

The purpose of this section is to propose hierarchies (or taxonomies or views) on resources usable when browsing the CLARIN repository. We build as far as possible on the insights obtained by investigating the abovementionded initiatives and registries. The browsing facility will be built on a federated search, i.e. once a resource is selected, the user will be pointed to the place where this resource, or service by which the resource can be accessed, is located. In other words, the CLARIN project will implement a federation of distributed language resources and tools.

On the top level, we distinguish between tools, lexical resources and corpora. The leftmost keywords are the top nodes of the category hierarchies or facets. Users can start from each of these top nodes and drill into each hierarchy arbitrarily deep, combining as many facets in their browsing as they like.

3.1 Views and facets for tools

```
Language
        Language Independent
        (unfolds a tree of language families,
         and per language family,
         a subtree of members)
Modality
        Spoken
        Written
        Multimodal
        Domain
        Segment-level
                Phoneme
                Grapheme
```

[9] However, research by Marti Hearst shows that a graphical interface which is similar to the Windows explorer has several downsides. The navigation hierarchy can become very large if several category trees are unfolded. Secondly, the idea of a resource residing in different hierarchies does not match well the *folder* metaphor on which this kind of interface relies. Since interface design is not our main concern here, we will not go into further detail. Cf. [2].

- Morpheme
- Syllable
- Word-level
  - Full-form
  - Lemma
  - Parts of speech
- Multiword
  - Collocations
  - Compounds
  - Phrase-level
- Sentence level
- Dialogue Turn level
- Paragraph-level
- Text-level

Approach
- Discrete
- Finite-state
- Context-free
- Statistical
- ML
  - Supervised
  - Unsupervised

Corpus tools
- Corpus editing tools

XML editors
- General editing tools

Indexing tools

Concordancers

Lexical tools
- Lexical acquisition
- Maintenance of lexical resources

Tasks
- Named Entity Recognition
- Speech synthesis
- Speech recognition
- Machine Translation
- Information Retrieval
- Information Extraction
- Question-Answering
- Text mining
- Co-reference
- Generation
- Summarization
- Latent Semantic Analysis
- Alignment

```
                Text
                Word alignment
                Phrase alignment
                Sentence alignment
        Speech
                Text-to-Speech
        Linguistic Annotation/Querying
                Annotation tools
                Querying tools
        Evaluation/Training
        Conversion
```

3.2 Views and facets for lexical resources

```
Language
          (unfolds a tree of language families,
           and per language family,
           a subtree of members)
Media/Storage
        Print
        Machine readable
        Lexical database
        Knowledge base
        Files
User
        Human
        NLP software
        Other
Linguality
        Monolingual
        Bilingual
        Multilingual
Language Stage
        Current
        Historical
Variety
        General
        LSP
        Group Language
        Dialect
        Regional Variant
Base
        Corpus
        CardIndex
        Other
Descriptive Level
```

```
        Form-based
        Content-based
Lexical Object
        Lexical Unit
        Synset
        Multi Word Expressions
        Lexeme
```

3.3 Views and facets for corpora

```
Language(s)
        (unfolds a tree of language families,
         and per language family,
         a subtree of members)
Language Stage
        Current
        Historical
        Mixed
Variety
        General
        Specific
                Technical (LSP)
                Sociolect
                Dialect
Linguality
        Monolingual
        Bilingual
                Aligned
                Non-aligned
        Multilingual
                Aligned
                Non-aligned
Presentation
        Aligned
        Non-aligned
Proficiency
        (Near)Native speaker
        Learner
Base
        Written text
        Printed Resources
        Born digital
        Spoken text
                Automatic transcription
                Manual transcription
        Multimodal Data
```

Writing System
Standard Orthography (version)
	Transliteration
	Transcription Scheme (IPA, SAMPA)
Media/Storage
	Print
	Machine readable Database
	Files

(All remaining views do only apply for machine readable corpora)
Encoding
Character set
		ASCII
ISO-8859-x
UTF-n
Other
User
	Human
	NLP software
	Other
Mode
Raw text
Annotated text
Annotation Scheme
	Pre-SGML/XML
	SGML/XML
		TEI
		CES/XCES
Other
Annotation Procedure
	Automatic
	Manual
Number of Annotation Levels
	Single
	Multi
Annotation Levels
	Linguistic annotations (treebanks)
		Morphology
		Parts of Speech
		Syntax
		Semantics
		Topological Fields
		Co-Reference
Other
	Non-Linguistic annotations

```
Document-Structure
Visual structure
```

4 Conclusions

The purpose of this paper was to present a motivation for the use of a collection of taxonomic trees as a means of access to the expected wealth of tools and resources which will be available through the CLARIN infrastructure. Browsing taxonomically organized views provides one way of finding the resources without having a precise knowledge about what is available in the repository and how it is organized. Another way to access a repository is through querying, but using this mode the researcher needs to have an idea about available descriptive categories and their values. CLARIN will have to serve both kinds of users: the casual user as well as the expert user. We therefore believe that both means of accessing the repository are necessary to offer. The taxonomies which we have outlined, and which are informed by other existing registries which are comparable to ours, have to be built on the metadata which will be provided by those people and institutions who will hold the data and want to make them visible and accessible via the CLARIN infrastructure. A view on faceted search can therefore inform the collection of metadata components, which is another major task of the CLARIN project. We have already made available an ad-hoc repository with around 600 resources through the CLARIN website (`www.clarin.eu`).

Note that on the conceptual level which is outlined in this paper, no decision has been made as to the representation language to which the taxonomical hierarchies should be bound. Topic Maps or RDF are possible candidates for such a binding, but this is an aspect of the implementation. The choice of taxonomical hierarchies is nevertheless a necessary precondition for the implementation of faceted browsing (cf. [1], pp. 98ff). An interface for faceted search will be implemented within the CLARIN project.

References

1. Foulonneau, Muriel / Riley, Jenn. *Metadata for Digital Resources.* Chandos Publishing 2008
2. Hearst, Marti. *Design Recommendations for Hierarchical Faceted Search Interfaces.* ACM SIGIR Workshop on Faceted Search, August, 2006
3. Max-Plank Institute Psycholinguistics. *Data Category Registry. DCR2 Requirements.* (ISO/TC 37/SC 4 N348) `http://www.tc37sc4.org/new_doc/ISO_TC37_SC4_N348_DCRregist_requirements_0%5B1%5D.2_EN.pdf`
4. Wittenburg, P. and Wright, S.A.. *ISO TC37/SC4 Infrastructure Note on Registry Databases.* `www.tc37sc4.org/new_doc/iso_tc37_sc4_N436`.