Language Resources, Taxonomies and Metadata (WP5 - White Paper)

Erhard Hinrichs, Lothar Lemnitzer, Andreas Witt

Tübingen University

Clarin Working Work Package 5: Languages Resources

1. Introduction

The purpose of this paper is to prepare a broad and detailed survey of language resources and tools (WP5; task C2) as one of the bases for the integration of these resources and tools into the emerging web service infrastructure (WP5; task R3). This paper is related to the efforts undertaken in WP2 on the development of the Clarin registry infrastructure. It therefore tackles the common WP5/WP2 from the WP5 perspective.

The purpose of a registry infrastructure is to provide information about and access to available language resources and services in a systematic fashion. The development of such a registry, 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 reflect the information and research needs of the potential users of the CLARIN infrastructure. This naturally leads to the idea of providing different views of the available tools and resources along different dimensions. Such views are likely to include the different languages, modalities of language (spoken, written, multimodal), intended user groups (e.g. linguists, historians, social scientists) and usages (e.g. information retrieval, machine translation, and speech recognition).

WP5 will undertake a detailed survey of existing tools and resources that can potentially be integrated into the CLARIN infrastructure. This will entail the collection of metadata associated with each resource. 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 IMDI.

All of this work will be informed by existing efforts and initiatives for cataloguing and categorizing languages tools and resources. Therefore, the present paper will first review prominent examples of existing registries (DFKI, ACL) and of existing catalogues (ELRA) as well as the IMDI metadata browser developed by the MPI. Informed by these initiatives, the second part of the paper 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[1]
- Max Planck Institute for Psycholinguistics: Data Category Registry DCR2 Requirements.[2]

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.

This document is meant as a basis for discussions in working groups 5 and for an exchange with other working groups. It will be complemented by other Clarin White Papers.

2. What registries 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 over existing methods, tools, resource 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 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 registry developed within Clarin. The Clarin registry must also support the search of resources by non-human agents, e.g. automatic processes.

One feature that all the registries we investigated share is that they do either 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 registry we are going to establish will not only allow the structuring of the resources along multiple views. It should be able for users to browse the repository following several path along different sequences of these views.

2.1 DFKI software registry

The Natural Language Software Registry (NLSR) was developed and set up by the German Research Institute on Artificial Intelligence (DFKI[3]), which is also a Clarin partner. The DFKI registry presents

[1] http://www.tc37sc4.org/new_doc/iso_tc37_sc4_N436_ontology_memo_peter_Sue_busan2007.pdf

[2] http://www.tc37sc4.org/new_doc/ISO_TC37_SC4_N348_DCRregist_requirements_0%5B1%5D.2_EN.pdf

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 it does this only to a certain extent. The NSLR does not consider it to be in their focus to present an exhaustive overview over resources. The users who are interested in these resources are referred to other institutions, especially the ELRA (see section 2.3).

The URL of the DFKI registry is http://registry.dfki.de/. On the left-hand side of the start and on all the other pages a navigation panel is shown.

It allows the user to navigate to several areas. The user can submit new resources (item 4 in the navigation panel) or to query resources (item 3). The taxonomical structure is accessible through item 2.

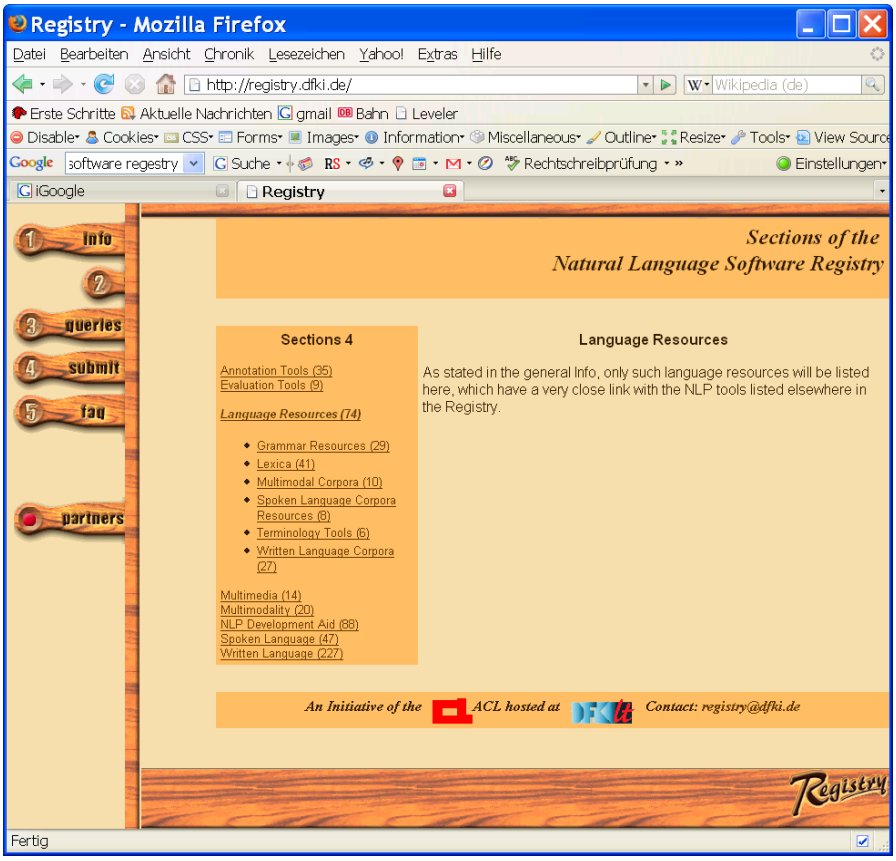

Fig. 1: The DFKI software registry

[3] The DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz) is locatated in Kaiserslautern, Saarbrücken, Bremen, and Berlin. The Institute for Language Technology which hosts the NLSR is based in Saarbrücken.

The tools are grouped into eight categories, e.g. tools for (manually) annotating resources, tools for processing of spoken or written languages, 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 group '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:

- Facial Movement & Speech Recognition
- Facial Movement & Speech Synthesis
- Speech and Gesture
- Text and Images

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:

- Language Detection
- Signal Analysis
- 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:

- Alignment Tools
- Comparative Linguistics
- Controlled Language Applications
- Corpus Analysis
- Deep Generation Applications
- Deep Syntactic Analysis
- Document Image Analysis
- Dynamic Hyperlinking
- Grammar Checking Applications
- Information Extraction Applications
- Information Retrieval Applications
- Language Guesser
- Lemmatizer
- Lexicon Management
- Morphological Analysis
- Morphological Generation
- Named Entities Detection
- Optical Character Recognition (OCR)
- Part-of-Speech Tagging Tools
- Partial Parsing
- Processing Mark-Up Languages
- Question/Answering
- Segmenter
- Semantic & Pragmatic Analysis
- Shallow Generation Applications
- Shallow Parsing Applications
- Spell Checkers
- Stemmer
- Summarisation Systems
- Terminology Extraction

- Terminology Management
- Text Classification
- Text Statistics
- Tokenization
- Translation Memory
- Written Dialog Environments
- Written Language Translation
- Written Language Understanding

When clicking on a category all the tools catalogued as a belonging to this group are listed. The number of the tools in a category is displayed in parenthesis directly after the name of the category. In general, tools are categorised as belonging to several groups.

2.2 ACL Web registry

The ACL web registry is a wiki-system provided and hosted by the Association for Computational Linguistics, one of the most important societies of the world-wide CL community. In contrast to the DFKI registry which has been established and is 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 registry.

The structure of the ACL resource wiki was last modified in April 2008. A screen shot of the registry's top level taken on April 22 lists the resource types 'Corpora', 'Knowledge Collections and data sets', 'Tools and Software', 'Dictionaries', 'Lexicons', and 'Language Resources'. Beside these categories there was also a second taxonomy "List of resources by language" and a container for "Uncategorized Resources". The following screenshot depicts the old structure.

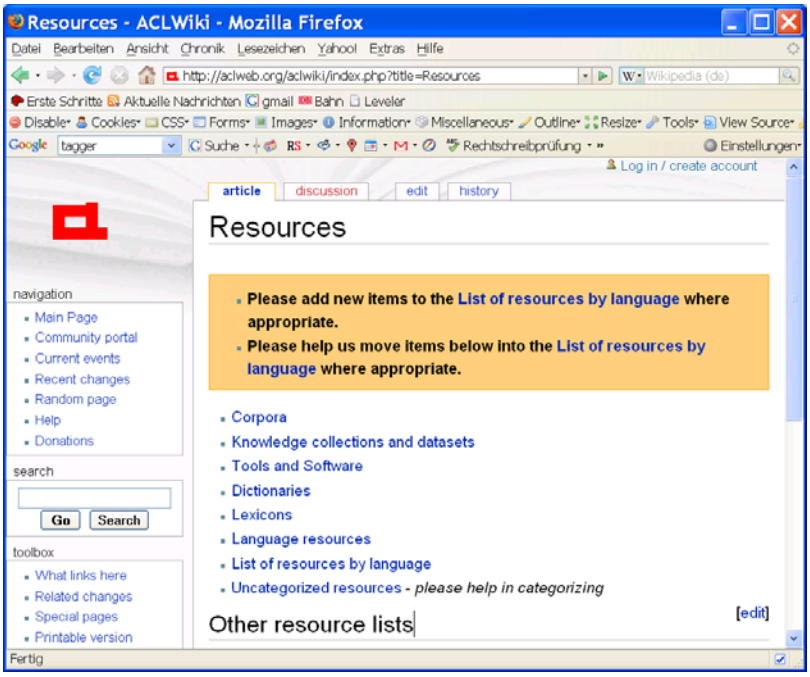

Fig. 2: The former structure of ACL web registry

The old structure allowed for listing and accessing different resource types by language and it was possible to switch from the listing of all the instances of one resource type to a listing of only those instances for a certain language.

The current version of the ACL only allows the user to browse the repository by language. The default language is English; other language resources are accessible via the link “List of resources by language”.

The resources are listed according to different categories. The categorisation differs between languages. The English resources, e.g., are grouped at top level into ‘Corpora’, ‘Dictionaries’, ‘Geographical words’, ‘Knowledge collections and datasets’, ‘Lexicons’, ‘Subject specific resources’, ‘Tools and Software’, and ‘Uncategorized resources’, whereas the top level categories for German are ‘Corpora’, ‘Evaluation datasets’, ‘Lexicons’, ‘Resource Access’, and ‘Timeline Analysis’ and the ones for Greek are ‘Machine translation systems’, ‘Corpora’, ‘Named entity recognition’, ‘Natural language generation’, and ‘Bibliography’. The main reason for these differences lies in the de-central organisation of wikis. Users are free to define their categories as they like.

Fig. 3: The new structure of ACL web registry

Some of the ACL categories are further sub-structured. 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
- 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 ELRA

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 one would like to purchase, a catalogue is provided by ELRA. This catalogue is accessible via WWW frontend.

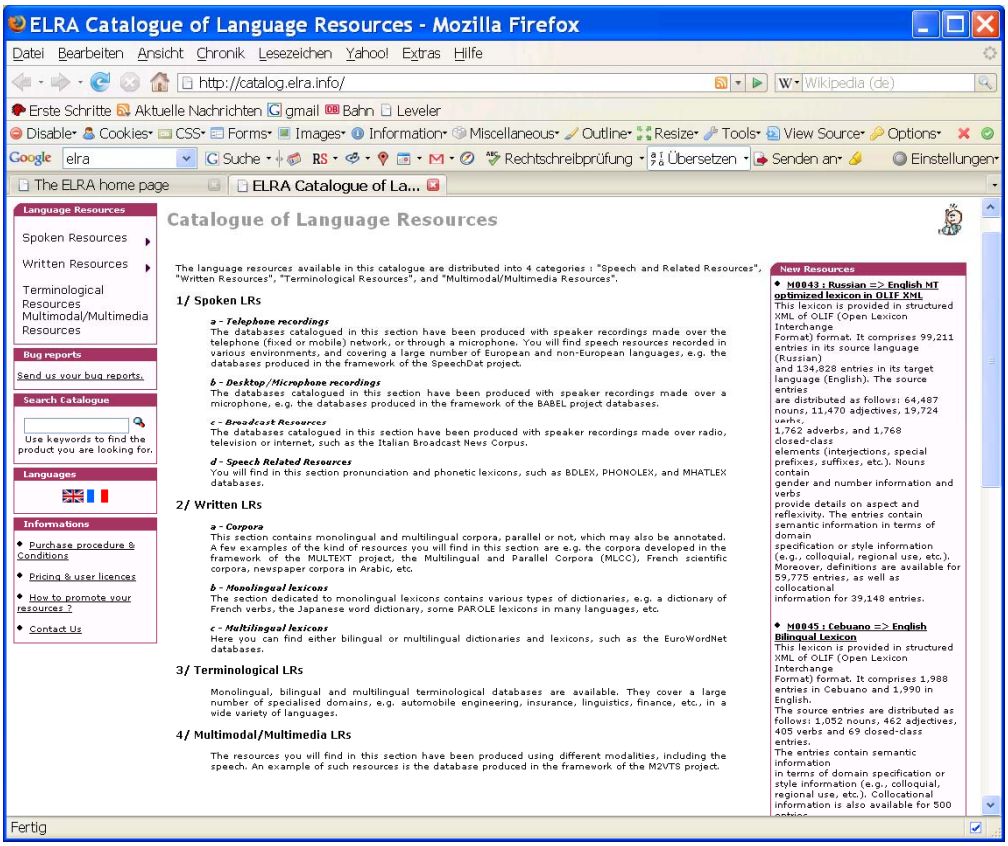

Fig. 4: Front page of the ELRA catalogue

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 LRs are not further subdivided. Written resources are grouped into the three categories: 'Corpora', 'Monolingual lexicons', and 'Multilingual lexicons'. ELRA's catalogue for spoken LRs substructures the resource type into 'Telephone recordings', 'Desktop/Microphone recordings', 'Broadcast Resources', and 'Speech Related Resources'. In the browsing mode, the ELRA

catalogue does not offer a “by language” feature. The search by language can be evoked by submitting a general query.

2.4 IMDI

Language resources using the IMDI metadata scheme could be accessed by 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 a hierarchical access to the resources. This can be done by organising the resources in the form of tree structures, consisting of nodes that group files together. Resources can be freely grouped by 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. A tool that supports carried out the creation of these hierarchies is the IMDI Tree Builder. It allows for creating trees based on IMDI metadata descriptors.

To display these structures the IMDI Tree Browser is used. The next figure shows a web access to IMDI data. The tree structure is displayed at the left-hand side of the editor.

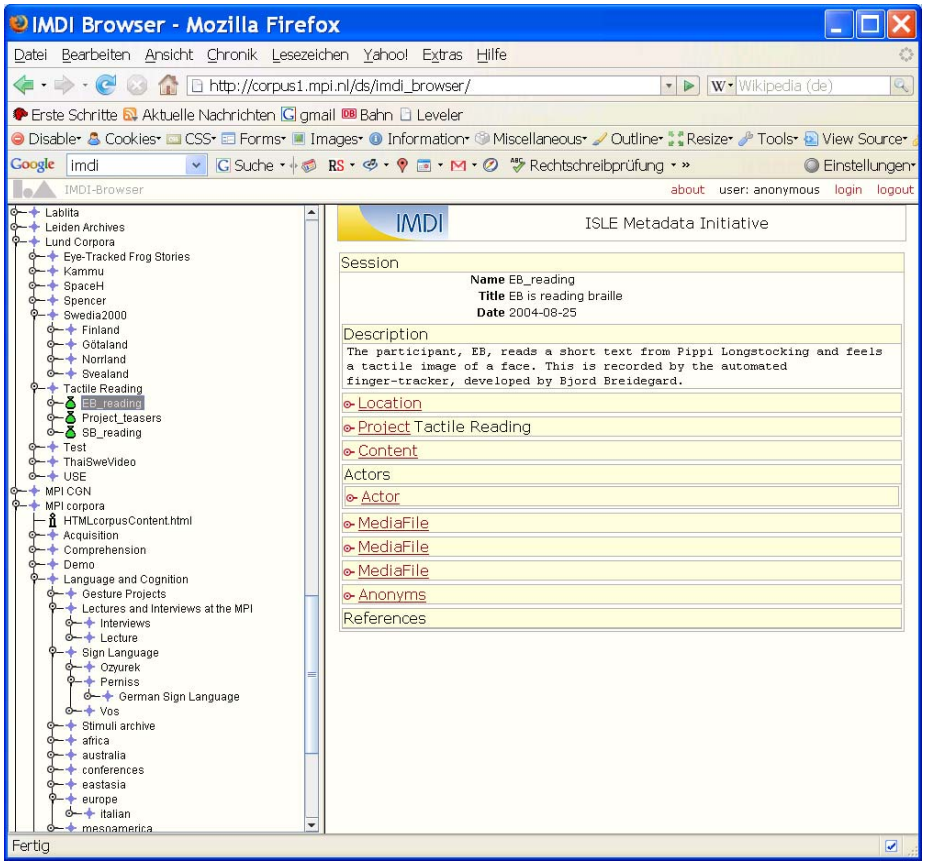

Fig. 5: The IMDI Browser

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 by the WP5 leader and his group, 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. These views should be discussed by the WP5 group members in order to refine them and to come to a shared agreement.

The IMDI browser could serve as a tool for retrieving tools and resources according to different predefined modes. Predefined hierarchies of the language resources and tools should established by the CLARIN project.

4. Possible Views

The purpose of this section is to propose hierarchies (or taxonomies or views) on resources usable when browsing the Clarin registry. We want to emphasise that the following pages are meant only as a first proposal that can serve as a basis of discussion by the three WP5 working groups. At the same time we have tried to build as much as possible on the insights obtained by investigating the initiatives described above.

4.1 Possible views on tools

These views should be discussed, elaborated and finally defined by Clarin Working Group 5.1.

- Language
  - The tree of language families
- Modality
  - Spoken
  - Written
  - Multimodal
- Domain
  - Segment-level
    - Phoneme
    - Grapheme
    - 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

4.2. Views for the lexical taxonomy

These views should be discussed, elaborated and finally defined by Clarin Working Group 5.2.

Definitions

Lexical resources summarises several resource type, especially dictionaries, lexicons, and terminologies.

Dictionary: a lexical resource containing information about linguistic aspect of the addressed lexical unit

Lexicon: a lexical resource containing information about the objects, SOAs etc which the addressed lexical unit signifies. An encyclopedia is a type of lexicon

Terminology: a lexical resource containing linguistic information of the terminological units and or information about the objects, SOAs etc which these lexical units signify

Language
- The tree of language families

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
- GroupLanguage
- Dialect
- RegionalVariant

Base
- Corpus
- CardIndex
- Other

DescriptiveLevel
- Form-based
- Content-based

LexicalObjects
- LexicalUnit
- Synset
- Multi Word Expressions
- Lexeme

4.3. Possible views for the taxonomy of corpora

These views should be discussed, elaborated and finally defined by Clarin Working Group 5.3.

- Language(s)
    - The tree of language families
    - The language(s) of the texts included in the corpus
- 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
    - Learner's corpora
- Base
    - Written text
        - Printed Resources
        - Born digital (does only apply for digital resources)
    - Spoken text
        - Automatic transcription (does only apply for digital resources)
        - 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 (e.g. ASCII, ISO-8859-x, UTF-n)
- User
    - Human
    - NLP software

    - Other
- Mode
    - raw text
    - annotated text

All remaining views do only apply for annotated corpora

- Annotation Scheme
    - Pre-SGML/XML
    - SGML/XML
        - TEI
        - CES/XCES, etc.
- Annotation Procedure
    - Automatic
    - Manual
- Number of Annotation Levels
    - Single
    - Multi
        - Annotation Levels
            - Linguistic annotations (treebanks)
                - Morphology
                - (only) Parts of Speech
                - Syntax
                - Semantics
                - Topological Fields
                - Co-Reference
                - ...
            - Non-Linguistic annotations
                - Document-Structure (e.g. chapters, paragraphs)
                - Visual structure (e.g. pagination)

5. Summary

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 which researchers need for their research purposes. Browsing is but one way to access these resources. Another way to access them is through querying. We believe that both means are necessary to offer, in response to user groups with different levels of knowledge and expertise and with different search habits and needs.

The taxonomies which we have outlined, and which are informed by other existing registries which are comparable to ours, have to match the metadata which will describe the individual resources. It is therefore vital for our project that we achieve a common understanding of and agreement on the taxonomic categories before we start collecting descriptive data about existing resources.