Project ref. no.	IST-1999-10647
Project title	ISLE MetaData Initiative

Deliverable status	Public
Contractual date of delivery	October 2000
Actual date of delivery	23.2.2001 (subcontract was signed in September, the overview was published in the web in October)
Deliverable number	D10.1
Deliverable title	EAGLES / ISLE Overview of Metadata Initiatives and Corpus Metadata in Language Engineering and Linguistics
Type	Report
Status & version	Final version – the overview will be extended when there are new MD sets
Number of pages	18
WP contributing to the deliverable	WP10
WP / Task responsible	Peter Wittenburg
Author(s)	Daan Broeder, Freddy Offenga, Peter Wittenburg
EC Project Officer	Brian Macklin
Keywords	Metadata, language resources, language engineering
Abstract (for dissemination)	The ISLE Metadata Initiative which is part of the ISLE project has analyzed which kind of metadata elements or so-called header descriptions are used until now for describing language resources. All projects which were known to us and which were mentioned by the members of the steering and advisory boards of the Metadata Initiative were included in the overview. Until now every project defined their own set of descriptors, often no such descriptors were used at all. On the one hand initiatives such as Text Encoding Initiative (TEI) and Corpus Encoding Standard (CES) have defined very rich sets to describe many details of written corpora. On the other hand initiatives such as Dublin Core (DC) have defined a very limited set to describe web-based resources such that they can easily be retrieved. As far as we know no project has tackled the special requirements of multimedia/multimodal language resources so far.

EAGLES / ISLE

Overview of Metadata Initiatives and Corpus Metadata in Language Engineering and Linguistics

November 2000

Background

The ISLE Metadata Initiative (IMDI) is part of the ISLE project (WP10). The work was started by presenting a White Paper and organizing a LREC pre-conference workshop and forming a formal framework for the project that exists out of a Steering Board, an Advisory Board, and a network of interested persons from the Language Resource community. The first Steering Board meeting was organized in Athens at the LREC conference. After these events the institutes participating in the IMDI work started to create a broad overview about existing metadata proposals and header information sets. The metadata descriptions and catalogue characteristics used in many projects and institutions were analyzed and compared. Based on this overview a proposal for a metadata set for the "multi-modal language resource community" was worked out and distributed describing resources at the session level. This proposal was presented at conferences and meetings and discussed with specialists from related communities such as the Dublin Core and the MPEG7 communities. First proposals for metadata element sets to describe lexica and catalog or higher nodes in the metadata hierarchy have been worked out.

For details the official web-site can be used: www.mpi.nl/ISLE.

Acknowledgements:
We like to thank Laila Dybkjaer and Malene Wagner for their comments on this report.

Daan Broeder, Freddy Offenga, Peter Wittenburg
Max Planck Institute, Nijmegen

NOTE: The overview mentioned under 2. and 3. were published on the web-site http://www.mpi.nl/ISLE . They are appended to this document. The format was optimized for the web-based presentation, therefore we would like to apologize for the type of presentation in this report.

1 Introduction

Language Resources (LR) are collections of data representing examples of language being used, either directly, as in corpora, or as derived data, as in lexicons and ontologies. Fundamental and applied linguistic research has a long history of generating and using text-based language resources, and more recently multi-media language resources have been exploited in linguistics and related areas such as sign language, anthropology, computer linguistics, artificial intelligence, phonetics, psychology, speech recognition, multi-modal research and man-machine interface design. They are used for a variety of purposes. Linguists use them to create and test new linguistic hypotheses; speech recognition engineers use them to test speech recognition devices and to set recognition parameters. Increasing amounts of money are being spent on creating new language resources and extending language resources to combine a variety of inputs (sound, video, eye tracking,…) and to incorporate multi-modal annotations.

People have always referred to such language resources in terms of basic global characteristics such as "the resource includes speech by a 6 year old male Tamil speaker born in a farming environment" or "this resource includes pointing gestures and speech utterances recorded when people were asked for directions to the railway station". We call this kind of data, which briefly summarises or characterises the content of the resource, its metadata-description. Most language resources include this data in "headers". These are either part of the resource itself or exist as separate files in a corpus-specific format. The examined projects define their own header structure and content, appropriate to the goals of the project. Special tools or simple ASCII editors were used to display and/or access the metadata.

The development of the World Wide Web, with its linked web pages, provides new opportunities. We envisage a universe of linked metadata-descriptions offering the interested community information about existing language resources. This universe should be accessible via the Internet with appropriate tools for browsing and searching. Such a system could save researchers and industry a lot of time in locating the appropriate LRs.

We are sure that the general addition of Metadata-Descriptions to Web-accessible data will eventually revolutionise the way the Internet is used. Thus the language resource community should test these new mechanisms at an early stage, both to confirm that the mechanisms and procedures now becoming available are capable of providing the services required by the language resources community, and to select the specific mechanisms and procedures which best match the needs of the community.

Making a universe of linked metadata-descriptions available for browse and search operations (resource discovery) requires the development of a standard for the structure and semantics of these metadata-descriptions. But language resources vary considerably and there is a heterogeneous user community, so we have to ask ourselves whether the requirements for handling such a variety of applications can be captured within a single standard. We have to look at the ways other communities have handled similar problems. The Dublin Core standard defined by the librarian's community would seem to be a relevant example. If we can exploit the experience built up in the development of this and other evolving standards, it should be possible to propose an ISLE-Metadata-Description standard within two years.

1.1 Previous Work in the Language Resource Community

Both the Bird/Liberman annotation web page and the MATE project list a large number of annotation schemes and projects. Of the 54 annotation schemes and projects now listed on the annotation web page, some are of only academic interest. The MATE deliverable D3.1 reviews eleven annotation tools in some detail. None of these reviews discusses how the various annotation schemes handle "header" data as opposed to "annotation" data. MATE deliverable 1.2 contains considerations of which header file information may be needed but does not go into much detail.

The CHILDLES project developed the CHAT format in the mid-1980s. This includes header data within what are essentially text files. A great many people have adopted this format and large corpora of child speech exist, but the format has its problems. The Text Encoding Initiative, which ran from 1990 to 1994

(in the first instance), solved some of these problems by adopting the SGML mark-up language, but continued to embed header data within the annotation files. The TEI initiative was influential, as can be seen from the Corpus Encoding Standard (CES) and MATE introductory web-pages, but does not constitute an encoding scheme - for instance the contemporary British National Corpus format, the Corpus Data Interchange Format (CDIF), though strongly influenced by TEI, differ in important details. The BBAW (Berlin-Brandenburgische Akademie der Wissenschaften) digital dictionary of German, though TEI-compliant, adds specialised annotation and separate index files aimed at supporting rapid search by the Alta Vista search engine. The index files include pointers to header metadata.

TEI, like CHAT, was text-oriented, but its adoption of the extensible SGML format proved to be a crucial and influential stage in the evolution of the modern XML-based, multi-media oriented formats such as MATE, LACITO and ATLAS, which include provision to reference annotation layers back to the raw audio and video (if available). In general, the information encoded in the older formats can be organised into annotation graphs, and machine-recoded into more modern formats, and both CES and MATE now restrict themselves to the XML-compatible sub-set of SGML.

At the Max Plank Institute for Psycholinguistics, the practical demands of speech and language research have required the development of several efficient multi-format tools. The EUDICO tool operates on a variety of formats, including CHAT, Shoebox and Tipster. The Spoken Childes Tool looks for CHAT formatted files, but can be used on European Science Foundation Second Language (ESF) files which have been recoded into the CHAT format. It is expected that the Bavarian Archive of Speech Signals (BAS) Partitur format could be recoded in a similar fashion.

In principle these tools can access some multi-media files over the Web. A multi-media file held in a web-accessible electronic archive is not necessarily web-accessible - the Australian AIATSIS electronic archive (Australian Institute of Aboriginal and Torres Strait Islander Studies) apparently includes multimedia documents, though none of them seem to be actually Web-accessible.

The University of Helsinki Language Corpus Server offers log-in access to computer corpora of more than 50 languages, but in so far as the corpora have been tagged, the tags differs from language to language, and the tags are not identified by either SGML- or XML - markup symbols (though the tag set used to label the Uralic language corpora is available on the web-site).

The web-pages that point to Web-accessible multi-media files can include some header data. It is effectively impossible to search the universe of web-pages for these specific pages on the basis of the header data they contain.

Since 1998 the Max Planck Institute for Psycholinguistics has been active in the development of annotation standards, and its Browsable Corpus scheme now being developed envisages the conversion of arbitrarily structured non standard header metadata of language resources files into uniformly tagged metadata, held separately in machine-searchable web-pages, to form the basis of a browsable and searchable universe of metadata resource descriptions.

1.2 Requirements of the LR Community

We have to produce work definitions of the terms "language resource" and "language resource community" for this metadata project.

"Language resources" are data which primarily document communicative acts of humans in some form of recording and/or descriptions, both directly as in corpora, or at higher levels of abstraction as in lexicons and ontologies. Similar resources have been created, by researchers working with chimpanzees where the communicative acts are studied and annotated. This work doesn't involve language as it is usually defined so it lies outside the scope of this work, but there should be enough flexibility (vocabulary, structure) such that this sort of work could be included. Human communication is either verbal or non-verbal (gesture, facial expressions, etc.); i.e. the basic material is either an audio recording or a video recording including audio tracks. When recording dialogue or discussion we can be confronted with several audio and video

tracks. We have to make allowance for situations where other data is recorded by non-audio/video techniques such as eye- or body-movement, or where brain images, EEG (Electro EncephaloGram) signals or articulation data are also recorded.

These basic recordings (time series) are supplemented by various annotations. These are tiers of manually or automatically generated textual descriptions. The manually generated tiers can be either free text or code generated from constrained input sets. These annotations can be made for a wide variety of purposes, largely dependent on the needs of the research or engineering disciplines targeted at. Linguists usually add layers of orthographic transcription, English translation, morphological coding and syntactical coding, and may abstract higher level descriptions like lexicons and ontologies. In disciplines where body movements in dialogues are studied, many tiers of annotation describing the gestures of the subjects will be added. Traditional language resources were restricted to textual descriptions, originally because they preceded reliable portable sound recorders, and more recently because it was too expensive to generate digitized versions of the original recording material. Some modern language resources include only textual material because they document written language usage, or concentrate on higher level abstractions such as words or sentences, as in lexicons and ontologies, amongst other secondary resources.

The "language resource community" to be addressed by this document, are the groups of researchers and developers working with such language resources. These resources can be used by developers and researchers for a wide range of purposes. For example researchers will use such resources for theorizing or testing new hypotheses, or technology developers use such resources to train their statistical recognition machinery. It is not assumed that the language resource community is a monolithic whole, where everybody has the same interests. We can easily identify such sub-communities as anthropologists who would like to be able to structure their data-bases around geographic references which are of little or no interest to other sub-communities. There will be other sub-communities with equally specific interests.

Granting that this is an adequate definition of language resources and the language resources community, we want to make a first attempt at the requirements for metadata descriptions. We assume that the community is interested in being easily able to find out whether resources with certain characteristics are available, and what is required to access these resources. We also asume that some (sub-)communities have to be careful with allowing access to Language Resources (LR) and will be interested in mechanisms that allow only controlled access to their LRs.

When looking for usable resources the users are not interested in a time consuming analysis of their contents, but they are interested in metadata which describes the form and content. A researcher working on longitudinal studies for Nordic languages might be interested in finding those resources which contain recordings of Scandinavian children taken every two years during the first 15 years of life. A producer of machinery which has to detect pointing gestures automatically, might be interested in finding those resources which contain video recordings where gestures are recorded and pointing is coded.

A possible solution is a universe of linked metadata descriptions of such resources which contain the relevant descriptors and pointers to the resources and to the institution that owns them. What we also need is a web portal, which can open this universe to the interested user. Finding the relevant descriptors via the Internet can either be done by browsing the linked metadata-descriptions, or by formulating and executing queries directed at a search engine, or by applying a mix of those two strategies.

Efficient browsing is dependent on the availability of intuitively understandable hierarchies formed by grouping resources together that share certain (metadata) characteristics. Experience shows that such hierarchies are difficult if not impossible to establish in large and anonymous groups. Nevertheless, it is useful to be able to navigate visual representations of such search spaces. If the domain is not too broad, search strategies which can operate on exact assertions can be more powerful. Efficient searching requires a data-space with very well defined categories, with precisely defined attributes which are restricted to a bounded range of values, which means that the metadata descriptions should be based on well-defined assertions, expressed in a common syntax.

Another requirement for efficient searching is, of course, a wide agreement about the categories that describe the relevant aspects of the resources for the whole community. The vocabulary has to be specified

and the semantics have to be laid down in accessible documents. Finding such categories and specifying their semantics is a time-consuming operation in heterogeneous communities. Having a too restricted vocabulary implies that certain fine-grained distinctions will not be available in the metadata description universe, which in turn requires that enough flexibility has to be available for sub-communities to add specific descriptions they might find important. This document will not go into further detail about the metadata vocabulary to be used. To chose the vocabulary a network of interested people has to be formed and a structure of discussion has to be described. In chapter 2 we will refer to the descriptors used in the traditional headers and to the metadata description format for the Max-Planck-Institute for Psycholinguistics. What can be said at this moment is that for the EAGLES/ISLE initiative the only feasible approach is one where we restrict ourselves to describing "language resources" as required by the language resources community.

It would be advantageous to maintain compatibility with other metadata-description mechanisms used on the World Wide Web. XML is the accepted standard for the syntax of metadata-descriptions and Resource Description Framework (RDF) is seen as possibly providing a framework for specifying structure and semantic relations to which the LR community could conform. This is described in more detail in the following chapter. We have to make sure that metadata which is already available in the web - such as addresses of people or companies - can be re-used; i.e. intuitively understandable concepts should be shared with other communities, if the chosen metadata mechanisms are compatible.

1.3 Work in W3C and related communities

The enormous increase in the number of web pages and the seemingly endless variety of information available on the Internet has made it necessary to think about new access strategies. Metadata-descriptions are seen as a means to provide searchable spaces. According to T. Berners-Lee, one of the driving forces behind the World Wide Web, metadata is machine understandable information about web-resources. The architecture of metadata is represented as a set of independent assertions. The Platform for Internet Content Selection (PICS) initiative was the first to use metadata to allow parental supervision of the web-data children can access. Companies such as Microsoft and Netscape identified similar needs and came up with the WebCollections and MetaContentFramework proposals respectively. Librarians as a community followed by defining the Dublin-Core standard, which can be used to describe contents of Digital Libraries in the most general sense.

The Dublin-Core (DC) defines 15 core elements and describes their meaning in web-accessible documents. This limited set of core elements can be refined by qualifiers to describe specialised documents. The core elements contain specifiers for topics such as the creator or title of a document, or the language it is written in. XML was chosen as the formalism for the syntax of DC and for structuring the metadata documents. The core set was limited to 15 elements to achieve semantic interoperability and give general users a uniform and easy-to-understand description. The definition of the vocabulary and the semantics is still going on and seems to be consuming a great deal of work.

At the moment it is still debated whether DC qualifiers can only be introduced by refining the semantics of the corresponding element or whether it is acceptable to extend the semantics. A severe problem here is that the semantics of the DC elements are not that precisely defined. Since DC now often is seen as a metadata umbrella for many different communities, these weak definitions could turn out to become a problem in the long run.

This plethora of initiatives lead the World Wide Web Consortium (W3C) to define a general framework for metadata descriptions called the Resource Description Framework (RDF). The RDF initiative is a very interesting one since it offers a unifying framework such that several different communities can share metadata definitions. For example names of persons will be used in many communities for different purposes. In the LR community a name could define a person who created the syntax annotation of a certain corpus file. Searchers might be interested in the affiliation of this person. The LR community could specify their own coding for the affiliation of such a person or they might chose to rely on the work of the electronic business card (vCard) community, since the vCard community's metadata descriptions associate

persons with a number of characteristics which include affiliation. The RDF standard makes it possible to combine the metadata vocabularies of various communities.

RDF uses the standard XML syntax. The vocabularies and the semantics are described by the various communities and RDF just offers a structural framework to bring these together. Name Spaces can be defined to refer to the definitions of various communities.

The MPEG7 standard initiative uses the term "metadata description" as already mentioned. MPEG7 is to define a standard description of the structure and content of movies such that people can search them to find specific scenes and such that the MPEG decoder can process the metadata stream in real time to do selections etc. MPEG7 therefore views all descriptions which are associated with a movie as metadata. This includes elements which describe the whole movie as well as elements which describe a certain feature of a sequence of video frames. Since MPEG7 decided to use XML as the underlying syntax for their descriptions, it is not surprising that they have chosen this approach. XML allows the user to create a hierarchy of such descriptions, starting at the top with the whole film as a subject of description, and ending with descriptions of the individual frames where necessary.

Of course, it would be possible to integrate the elements describing the whole content into separate files and integrate them into a browsable universe.

1.4 Problems to be solved

This chapter is devoted to listing all the major problems, which have to be tackled and solved within the EAGLES/ISLE metadata initiative. Some of these problems are inter-dependent, and the ordering of the list reflects this, but other orderings are possible.

Goal

The goals of the EAGLES/ISLE metadata-initiative given above were originally set out in the proposals submitted to the EC, but they have to be accepted as appropriate by the language resources community before any significant work can be done.

There are two deliverables beyond creating an overview:

1. A proposed standard for metadata descriptions for Language Resources.

2. A showcase demonstrating what the proposed standard would look like and how it might be used.

There is no guarantee that the proposed standard would actually be used, but there is an obligation to publish the proposal within the language resources community. The MPI expects to provide a showcase demonstrating the use of the proposed standard.

Influences

The EAGLES/ISLE work on a metadata standard has to reflect two influences; on the one hand the expectations and demands of the language resource community, which has to specify its requirements, and on the other the current state of metadata initiatives, especially those arising within the World Wide Web Consortium (W3C). The proposed standard will have to map the requirements of the languages resources community, ideally onto the structures being built under the aegis of the World Wide Web.

Scope of LR and Community

As mentioned above, we have to define the range of the Language Resources to be covered by our metadata initiative. This effectively determines the data which has to be covered by the metadata-descriptions and the scope of the standard, and defines the community which we want to address.

We have to identify the sub-communities which may require the inclusion of specific - possibly unique - data in the metadata elements to be used to characterize their Language Resources. There are several dimensions within which sub-communities can be identified. When we look at the types of language resource we can distinguish textual corpora, annotated corpora, multi-media corpora, lexicons, typology databases, grammar notes, notes about sound-systems, ontologies, and others. This may be the best place to start. Another dimension can be defined by the specific research or development needs of particular sub-communities. Anthropologists will want to include other descriptors for multi-media/multi-modal corpora than will people designing man machine interfaces. While anthropologists are for example interested in knowing the educational background of their subjects, language engineers at this moment will be more interested for example in the quality of the recordings. However, it is known that actual wishes can change quickly dependent on the possibilities of technology. Until recently few engineers were interested in information about modalities included. Now, this is a very important information. Engineers looking for corpora which include gesture encoding would like to have a quick possibility to see where such resources are available.

Structure of Metadata-Descriptions

XML seems to be a natural choice for the underlying syntax of the metadata descriptions. It has the necessary power to express the required structure and also will allow us to use an expanding range of available software for parsing and generating purposes. Probably there will be a core set of mandatory metadata elements which will be appropriate for all types of language resources. The presence of the core set would identify a file of metadata-descriptions as describing a language resource, and allow a basic browse and search tool to navigate this universe of metadata descriptions. There would also be a significant range of common elements which would be optional, because creating metadata descriptions is a time-consuming task and we would not want to burden the users with inserting unnecessary data. In order to achieve the desired flexibility to cope with idiosyncratic sub-communities and specialised classes of LR, there would be open extensions which can accommodate specialised vocabularies of metadata elements.

The detailed discussion of the management of such a flexible system requires further study.

It is not as yet certain that the model proposed by RDF is suitable for our needs since there are problems with its property centricism and its difficulties in expressing structure and value restrictions. MPEG7 has apparently rejected it as incapable of handling multiple layers of annotation, but since we are not proposing to use RDF to encode annotation data, this doesn't seem to be a problem for us. Adopting existing standards has obvious advantages if they can provide the services required without importing too many unnecessary features, but there is no virtue in adopting an inappropriate standard.

Scope of Metadata

We have to define the scope of the metadata; i.e. what type of solutions do we want to offer. It is clear that we want elements, which describe form and content of the related LR. However, we need to decide if we also want to include e.g. elements

- which indicate the set of annotation tiers included
- which indicate standard or standards used to create the annotations (EAGLES …)
- which point to the original video tapes (which will probably never be used)
- which have some information relevant for certain tools
- which contain statements about rights, forms of accessibility, and payment
- Information pertaining to distribution, licensing and pricing of the LRs

Metadata Element Vocabulary

The definition of elements will take most of the time. The work will include bringing together specialists for the major dimensions to be identified (type of resources, sub-communities). Existing header file definitions and current practice at such institutions as the MPI have to be analyzed to identify descriptors which have already proven their usefulness. From these discussions the set of elements to be part of the standard proposal has to be defined. For this elements the semantics then have to be defined and published as a web-accessible document. Examples have to be created to make the semantics as explicit as possible to the normal users.

Metadata Element mapping

Some metadata elements are not to be directly accessible via the Internet. In many cases the names of subjects may not be made public and here a kind of mapping mechanism must be used to replace the true name of the subjects by an alias that can only be resolved by a mapping file which is not publicly available.

Re-usage of metadata element definitions from other communities

Assuming that a unifying mechanism such as RDF can be used, we have to check whether it makes sense to make use of elements already defined and used by other communities such as DC, vCard etc. These definitions have to be stable and must appear intuitive to the language community.

This sort of judgement seems to require access to the people who defined the elements for their various communities. The history of the Dublin Core initiative suggests that it can be difficult to just create definitions which mean the same thing to different members of the same community, let alone the members of different communities.

Requirements for tools

Ultimately we have to define requirements for tools to work on metadata descriptions.

There have to be metadata description **editors**, which help the common user to enter the descriptions. These editors have to support the user by making the semantic descriptions of the elements available on request, and have to be flexible enough to cope with the structural flexibility required of the metadata description format.

We need suitable **browsers**, which understand the structure of the linked metadata description files and provide graphic support for the user during navigation. Since the standard browsers such as Netscape and Internet Explorer do not give us the desired functionality, we can either enhance their functionality by developing special applets or develop new proprietary browsers specific for the LR metadata descriptions. We also need **search** tools, which can cope with the metadata-description file structure and any metadata elements taken over from other communities. The search tools have to use the links between the metadata-descriptions and knowledge about available metadata descriptions efficiently. In cases where there are many metadata-descriptions to be processed, optimization techniques such as caching pre-parsed metadata description files in a data-base can be required to keep response times within reasonable bounds.

Practicable Scenario

In our final documents which will be written at the end of the ISLE metadata project we should describe a practicable scenario. This includes such topics as determining

- where to store the metadata descriptions
- ways to register and link the metadata descriptions
- ways to build browsable hierarchies
- ways to supervise the linking of new descriptions to the existing universe

- the requirements for centers which could establish and maintain such a universe

The Internet is growing dramatically and we need to understand how we can use it. As yet, nobody has tackled the topics mentioned above. We are faced with new challenges and must devise new solutions. The proposals which should emerge from our EAGLES/ISLE project are intended to be signposts on the route to establishing the sort of web we need.

1.5 Recent Developments

The Ithaca Technical Committee meeting of the Open Archives Initiative (OAI) set a broader scope for the archives to be included in it, i.e. OAI left the narrow scope of focusing on library types of service. OAI now also includes metadata services in so far that a metadata harvesting protocol was developed and that it was decided that DC is the metadata set adopted for the OAI. This decision makes sense since many communities will make use of the OAI umbrella. The opening of OAI makes it an interesting candidate to act as an umbrella for many archives with widely differing contents.

Recently, the LDC presented the OLAC concept (Open Language Archive Community) which is based on the OAI Metadata set (which actually is the DC standard) and the OAI harvesting protocols. The DC metadata set will be slightly extended such that the language a resource is written in and the language a resource is about can be differentiated and can be used in queries to retrieve the corresponding resources. This top-down concept is seen as complementary to the bottom-up concept applied by the ISLE project. While in the former a relatively weakly specified and universal metadata set (DC) is taken to meet the needs of general services, the latter started to analyse the needs of the community to categorize their resources. Both approaches are essential and address different needs. As in many other cases harmonization efforts have to be carried out to map the emerging IMDI set to the Dublin Core set. The responsible persons from OLAC and IMDI have discussed this issue and will work on a solution such that IMDI metadata records can be harvested by OLAC based services.

2 Project Overviews

Overviews from relevant metadata initiatives and corpus metadata are given to get a clear picture of the metadata used by the language engineering community. Projects containing an apparent notion of corpus metadata were included in the overviews. The well-known metadata initiatives CES, DC and MPEG were also examined in order to see which elements overlap with elements from metadata from the language engineering community. We also looked at the catalog data from the resource agencies LDC and ELRA, since they also contain descriptive data. Each corpus overview includes a short description of the corpus / initiative. The structure of the corpus and the corpus itself is described and there is information about the documents contained in the corpus. Since most of the time the metadata is found in the headers of the documents, these headers are also described in the overview. The metadata elements and their definitions are given in the form of a table. Groups of elements are indicated by indentations in the table.

It has to be noticed that all relevant resources are included in this overview that we are aware of and that were indicated to us by members of the Steering and Advisory Boards. Those projects that are in the process of defining a metadata set, that used for example a set compliant with one of those mentioned in chapter 2 and 3, or that have a structured corpus unsupported by metadata descriptions are summarized in chapter 4. They don't add information to the overview.

In doing so as described we believe to give a survey of best practice and trends in metadata descriptions. However, it became clear that the idea of creating structured and web-accessible metadata descriptions is as fairly new one to the community. The overview will be expanded during the rest of the project when new initiatives and projects can be indicated.

2.1 Browsable Corpus (BC)

The Browsable Corpus concept was introduced at the Max Planck Institute for Psycholinguistics (MPI) to make resource discovery easier by defining metadata descriptions for language resources. The structure of linked metadata descriptions can be browsed and searched.

2.2 Corpus Encoding Standard (CES)

The Corpus Encoding Standard (CES) is an encoding standard for corpus-based work for use in the language engineering community. The CES is an application of SGML and conformant to the Text Encoding Initiative (TEI) guidelines but dedicated to corpora. Therefore, TEI was not added to this overview.

2.3 Codes for the Human Analysis of Transcripts (CHAT)

CHAT is the format used for the CHILDES (Child Language Data Exchange System) project.

2.4 Dublin Core (DC)

The Dublin Core is a metadata element set intended to facilitate discovery of electronic resources. Originally conceived for author-generated description of web resources, it has attracted the attention of formal resource description communities such as museums, libraries, government agencies, and commercial organizations.

2.5 European Language Resources Association Catalog (ELRA)

ELRA (European Language Resources Association) is an organization to promote the creation, verification, and distribution of language resources in Europe. The catalog includes a wide range of corpora including speech corpora, written corpora and terminology corpora. Here we focus on descriptors used in the catalog.

2.6 European Science Foundation Second Language Databank (ESFSLD)

The ESFSLD is a computerized archive of data collected by research groups of the ESF-project in five European countries: France, Germany, Great Britain, The Netherlands and Sweden. The project concentrates on the spontaneous second language acquisition of forty adult immigrant workers living in Western Europe, and their communication with native speakers in the respective host countries.

2.7 Gesture Databank (GDB)

The Gesture Database from the Max Planck Institute in Nijmegen consists of the video recordings of speech and gestures that spontaneously accompany speech, and the annotations regarding gesture and speech in the recording. The recordings were made in different cultures, including the Netherlands, Italy, the USA, Japan, Turkey, Australian Aboriginal communities, Mexico, Belize, and Ghana. Speech events are recorded that elicits spontaneous gestures, such as narration of traditional stories and autobiographical stories, description of the local environment, and route direction.

2.8 International Corpus of English (ICE)

The International Corpus of English began in 1990 with the primary aim of providing material for comparative studies of varieties of English throughout the world. Twenty centers around the world are preparing corpora of their own national or regional variety of English.

2.9 Linguistic Data Consortium Catalog (LDC)

The Linguistic Data Consortium supports language-related education, research and technology development by creating and sharing linguistic resources: data, tools and standards. The LDCs Catalog contains 168 corpora of language data. Again in this overview we focus on the descriptors used in the catalog.

2.10 Multimedia Content Description Interface (MPEG-7)

The Multimedia Content Description Interface (MPEG-7) is an ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) standard developed by MPEG (Moving Picture Experts Group). MPEG-7 aims to create a standard for describing the multimedia content data that will support some degree of interpretation of the information's meaning, which can be passed onto, or accessed by, a device or a computer code. MPEG-7 is not aimed at any one application in

particular; rather, the elements that MPEG-7 standardizes shall support as broad a range of applications as possible. Only some elements from the Multimedia Description Schemes (MDS) part of MPEG-7 which we think are relevant for language resources are described in the overview.

2.11 Spoken Dutch Corpus (CGN - Corpus Gesproken Nederlands)

The Spoken Dutch Corpus Project is aimed at the construction of a database of contemporary standard Dutch as spoken by adults in the Netherlands and Flanders. Upon completion, the corpus will contain approximately ten million words, two thirds of which originate from the Netherlands and one third from Flanders. The Spoken Dutch Corpus comprises a large number of samples of (recorded) spoken text. In all about 1,000 hours of speech.

3 Global Overview

As already mentioned we refer to the web-site (www.mpi.nl/ISLE/overview/overview_frame.html) for the actual overview and prefer to not include it in this document for mainly presentational reasons.

In the global overview a mapping is made between the metadata elements from different initiatives. Each column in the overview represents the elements specified by an initiative. In the first column definitions are given that agree with the definitions of the elements from different initiatives. Since the overview is made only to give an overview of the metadata used in the initiatives and not to find the exact differences between the metadata elements from the initiatives, small semantic differences between matching elements (from a row) are allowed. When the semantic difference is too large according to our knowledge, the deviating element is placed in another row and gets another definition.

The overview is split in two parts. The first part contains the elements which have a clear definition. The remaining elements in the second part of the overview are the elements without a clear definition because definitions are missing or because the given descriptions aren't specified well enough.

4 Other Projects

The following important initiatives and projects are acknowledged but not included in the overviews for different reasons. They are either looking for a metadata standard, are compliant with a set which is already included in the overview, or don't yet use structured descriptions in the form discussed in this note.

4.1 Archive of Indigenous Languages of Latin America (AILLA)

The AILLA is a project to develop a web-based archive of linguistic materials of the indigenous languages of Latin America.

4.2 Alaska Native Language Center (ANLC)

The ANLC is recognized as the major center in the United States for the study of Eskimo and Northern Athabaskan languages. It is the center for research and documentation of the twenty Native languages of Alaska. There is no ANLC overview because ANLC is in the process of getting their metadata in line with Dublin Core.

4.3 British National Corpus (BNC)

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. The reason why there is no BNC overview is because the BNC text encoding is TEI-conformant and we rely on the Corpus Encoding Standard (CES) claim that relevant elements for corpus encoding are selected from TEI.

4.4 Linguistic Data Archiving Project (LACITO)

The goals of the LACITO linguistic data archiving project are the conservation and the distribution of speech data. The maintainers of LACITO are currently looking for a metadata standard.

4.5 Michigan Corpus of Academic Spoken English (MICASE)

The on-line, searchable part of a collection of transcripts of academic speech events recorded at the University of Michigan. The maintainers give a clear overview of metadata descriptions in the form of speech events and speaker attributes.

4.6 Text Encoding Initiative (TEI)

The Text Encoding Initiative (TEI) is an international project to develop guidelines for the preparation and interchange of electronic texts for scholarly research, and to satisfy a broad range of uses by the language industries more generally. TEI is not included in the overviews because we rely on the fact that the Corpus Encoding Standard selected the relevant elements from TEI for corpus encoding.

4.7 University of Helsinki Language Corpus Server (UHLCS)

The University of Helsinki Language Corpus Server (UHLCS) is a multilingual corpus server located at the Department of General Linguistics, the University of Helsinki. The server contains computer corpora of more than 50 languages, including samples of minority languages and extensive corpora representing different text types. The UHLCS is currently structured as a language hierarchy. The definition of UHLCS metadata is in progress.

5 Conclusions

From the work on preparing the overview and developing the first draft IMDI metadata set which was carried out in partial overlap, we can draw the following conclusions:

1. The consequent usage of structured and web-accessible metadata descriptions is a fairly new area of work. The oldest attempt to define a set of descriptive elements for easy document retrieval is that of Dublin Core. DC originally was an enterprise of the librarians which started in October 1994. Later the perspective was broadened to include all web-accessible documents. To allow the various communities to accept DC as an umbrella for metadata descriptions it was decided to only define a very limited set of 15 elements and to specify their semantics only loosely.
2. In the area of metadata descriptions (and header data) for traditional language resources a large variety can be found. This ranges from the very limited and specialized catalog data of the leading language resource agencies such as ELRA and LDC to the very exhaustive and TEI-compliant set defined within the Corpus Encoding Standard. This was not designed to be a web-accessible set and not optimized for retrieval purposes.
3. Further we found a great interest in the metadata idea on most of the meetings we participated in. People want the kind of browsable & searchable structured web-space to more easily find language resources for their purpose. They also like the vision that browsing and searching leads to interesting resources and that they can directly start programs which work on them. This would solve all the platform and format dependency issues people have to live with now for many years and which costs so much money. Metadata descriptions are seen as being very useful in those cases where the resources are not openly available. MD descriptions at least would tell people very quickly which resources have been created.
4. Companies such as Lernhout&Hauspie are using databases with descriptions in house for exactly the same reason. They need to know which resources are available in the company, for what they can be used, and what their status is. Unfortunately they were not willing to make the structure of their database open.
5. For the interested people of the language resource community who gave responses the DC set was seen as much too limited to be useful. On the other hand a CES/TEI set was seen as too much devoted to textual material and too exhaustive for retrieval purposes.

6. Until now as far as we know there were almost no ideas of how to describe other types of language resources than corpora. Therefore, it was decided to first focus on a set for corpora and start thinking and discussing about a descriptor set for lexica.
7. For the IMDI project we took the conclusion that we have to develop a moderate set and where useful controlled vocabularies. It was also understood from the discussions that the members of the language resource community need a terminology that they directly can understand to be willing to create metadata descriptions. The consequence was that we decided to first derive a metadata set independent of the DC definitions and at second place provide a mapping between the coming IMDI set and DC.
8. It was also understood that developing a metadata set which has some chance of being broadly used will depend on the early availability of a demonstration and of tools. Therefore, it was decided within the IMDI project to slightly change the time schedule. It was planned to create a showcase to demonstrate a browsable hierarchy of language resources for the Official Opening Event of the European Year of the Language in Lund in February 2001 and to organize a major IMDI workshop in March 2001 in Nijmegen.

7 References

Bearman, D., et al (1999). A Common Model to Support Interoperable Metadata http://www.dlib.org/dlib/january99/bearman/01bearman.html

Brickley, D., Hunter, J., Lagoze, C. (1999). ABC : A Logical Model for Metadata Interoperability http://www.ilrt.bris.ac.uk/discovery/harmony/docs/abc/abc_draft.html

Broeder, D.G., Brugman, H., Russel, A., and Wittenburg, P. (2000). A Browsable Corpus: accessing linguistic resources the easy way. LREC 2000 Workshop, Athens

Feldweg, H. (1992). The European Science Foundation Second Language Databank

Heid, Evert, and Berman, (2000). Searchable Metaspaces - Draft for the EAGLES/ISLE workshop. LREC 2000 Workshop, Athens.

Hunter, J., James, D. The Application of an Event-Aware Metadata Model to an Online Oral History Project, http://archive.dstc.edu.au/RDU/staff/jane-hunter/OralHistory/paper.html

Hunter, J., Zhan, Z. (1999). An Indexing and Querying System for Online Images Based on the PNG Format and Embedded Metadata, http://archive.dstc.edu.au/RDU/staff/jane-hunter/PNG/paper.html

Ide, N. and Brew, C., (2000). Requirements, Tools, and Architectures for Annotated Corpora. LREC 2000 Workshop, Athens.

Lagoze, C. (2000). Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience, http://ncstrl.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR2000-1801

Lagoze, C., Hunter, J., Brickley, D. (2000). An Event-Aware Model for Metadata Interoperability, http://www.ncstrl.org/Dienst/UI/1.0/Display/ncstrl.cornell/TR2000-1800

MacWhinney, B. (1991). The Childes Project: Tools for Analyzing Talk

Oostdijk, N. (2000). Meta-Data in the Spoken Dutch Corpus Project. LREC 2000 Workshop, Athens.

Strömqvist, S. (2000). Optional extensions - a proposal for a flexible annotation system. LREC Workshop, Athens.

Suihkonen, P. (2000). On Meta-Descriptions for Cross-Linguistic Electronic Linguistic Data. LREC 2000 Workshop, Athens.

Wittenburg, P., Broeder, D., and Sloman, B. (2000). EAGLES/ISLE: A Proposal for a Meta Description Standard for Language Resources, White Paper. LREC 2000 Workshop, Athens

Wittenburg, P., Brugman, J., Broeder, D. (2000). Workshop on Annotation Architecture and Software Tools for Multi-Media Language Resources and Large Corpora, LREC Preconference workshop, Summary. LREC 2000 Workshop, Athens.

Wittenburg, P., Brugman, J., Broeder, D. (2000). Workshop on Meta-Descriptions for Multi-Media Language Resources. LREC Preconference Workshop Summary, Athens.

8 Web References

Aboriginal Studies Electronic Data Archive (ASEDA),
http://coombs.anu.edu.au/SpecialProj/ASEDA/fabout.htm

Alaska Native Language Center,
http://www.uaf.edu/anlc/index.html

Archive of the Indigenous Languages of Latin America,
http://uts.cc.utexas.edu/~ailla/index.html

Archives for Language Documentation and Description,
http://morph.ldc.upenn.edu/exploration/archives.html

The British National Corpus (BNC),
http://info.ox.ac.uk/bnc/

Browsable Corpus (BC),
http://www.mpi.nl/world/tg/lapp/browscorp/browscorp.html

Corpus Encoding Standard (CES),
http://www.cs.vassar.edu/CES/

Corpus Encoding Standard for XML (XCES),
http://www.cs.vassar.edu/XCES/

Dublin Core Metadata initiative,
http://purl.org/DC/

European Language Resource Association (ELRA),
http://www.icp.grenet.fr/ELRA/home.html

European Science Foundation (ESF) Second Language Database,
http://www.mpi.nl/world/tg/lapp/esf/esf.html

International Corpus of English (ICE-GB),
http://www.ucl.ac.uk/english-usage/ice-gb/

Linguistic Annotation,
http://www.ldc.upenn.edu/annotation/

Linguistic Data Archiving Project (LACITO),
http://195.83.92.32/presentation/index.html.en

Linguistic Data Consortium (LDC),
http://www.ldc.upenn.edu/

Michigan Corpus of Academic Spoken English (MICASE),
http://www.hti.umich.edu/m/micase/

MPEG Standards (MPEG-7),
http://www.cselt.it/mpeg/standards.htm

Multilevel Annotation, Tools Engineering (MATE),
http://mate.mip.ou.dk

Open Archives Initiative,
http://www.openarchives.org/

Open Language Archives Community,
http://www.language-archives.org/

Searching ICE-GB with ICECUP,
http://www.ucl.ac.uk/english-usage/ice-gb/icecup.htm

Spoken Dutch Corpus (CGN-project),
http://lands.let.kun.nl/cgn/ehome.htm

Talkbank,
http://www.talkbank.org/

Text Encoding Initiative (TEI),
http://www-tei.uic.edu/orgs/tei/

The Child Language Data Exchange System (CHILDES),
http://childes.psy.cmu.edu/

The Schemas Project, Forum for Metadata Schema Implementers,
http://www.schemas-forum.org/

University of Helsinki Language Corpus Server (UHLCS),
http://www.ling.helsinki.fi/uhlcs/index.html

Appendix : Workshop Organization/Participation and Presentations

Members of the IMDI working group within ISLE organized the following meetings where metadata descriptions were part of the topics. In fact the events and presentations were the greatest source of feedback we got to create the overview and to develop the first draft proposal for the IMDI set. The overview was available since October on the IMDI web-site and was subject of several discussions. Especially of interest was that communities in the US and related disciplines from the humanities such as philosophers and archeologists took notice of web-based overview.

- Wittenburg (MPI), Roy (U Odense) and Cunningham (U Sheffield) Wittenburg organized an international workshop about "Meta-Descriptions for Multimodal/Multimedia Language resources", which was held in Athens on May 29th at the LREC Conference. For details see www.mpi.nl/ISLE.

- Wittenburg (MPI) organized an international symposium about "Measurement Methodologies in Gesture and Sign Language", which was held during the Measurement Behavior 2000 Conference in Nijmegen on August 17th. For details see www.noldus.com/events/mb2000.

- Wittenburg and Brugman (MPI) organized the first DOBES Workshop about "Documenting Endangered Languages", which was held in Nijmegen on September 15/16th. For details see www.mpi.nl/DOBES.

- Wittenburg and Brugman (MPI) organized the second DOBES Workshop about "Documenting Endangered Languages", which was held in Hannover on January 12/13th. For details see www.mpi.nl/DOBES.

The following presentations about metadata issues were given:

- Wittenburg, P. 'Meta-Descriptions for Language Resources'. LREC Pre-Conference Workshop on "Meta-Descriptions for Multimodal/Multimedia Language resources". Athen, May.

- Broeder, D. 'A Browsable Corpus: Accessing Linguistic Resources the easy way'. LREC Pre-Conference Workshop on "Meta-Descriptions for Multimodal/Multimedia Language resources". Athens, May.

- Brugman, H., Wittenburg, P., Broeder, D. 'COREX – Corpus exploitation software for the Dutch Spoken Corpus'. Tilburg, December.

- Wittenburg, P., Brugman, H., Broeder, D., Russel, A. 'Infrastructure to support Gesture Research'. International Measurement Behavior 2000 Conference, Nijmegen, August.

- Brugman, H., Wittenburg, P. 'MPI tools and software architectures'. Talkbank Technical Workshop. Pittsburgh, October

- Broeder, D., Suihkonen, P., Wittenburg, P. 'Developing a Standard for Meta-Descriptions of Multimedia Language Resources'. Talkbank Workshop on Web-based Language Documentation & Description. Philadelphia, December.
- Wittenburg, Brugman, Broeder, D. 'Web-based Language Documentation &Description at the MPI and within DOBES'. Talkbank Workshop on Web-based Language Documentation & Description. Philadelphia, December.
- Wittenburg, P. 'Multimodality and Crosslingual Knowledgemanagement' EC Expert Meeting on "Emerging Issues in Crosslingual Knowledge Management". Brussels, June. Wittenburg, P. 'Meta-description standard for multi-media LR'. Dublin-Core Usage Workshop. Luxembourg, May.

- Wittenburg, P. 'Meta-Descriptions for Language Resources'. LREC Pre-Conference Workshop on "Meta-Descriptions for Multimodal/Multimedia Language resources". Athens, May.

- Broeder, D., Brugman, H., Russel, A., Skiba, R., Wittenburg, P. 'Towards a standard for Meta Descriptions for language resources'. LREC Pre-Conference Workshop on "Meta-Descriptions for Multimodal/Multimedia Language resources". Athens, May.

- Broeder, D., Suihkonen, P., Wittenburg, P. 'Developing a Standard for Meta-Descriptions of Multimedia Language Resources'. Talkbank Workshop on Web-based Language Documentation & Description. Philadelphia, December.
- Wittenburg, Brugman, Broeder, D. 'Web-based Language Documentation &Description at the MPI and within DOBES'. Talkbank Workshop on Web-based Language Documentation & Description. Philadelphia, December.

- Wittenburg,P. 'Meta-Beschreibungen im Internet'. 14th Computer Science Meeting in the Max-Planck Society. Göttingen, November.

Related Projects Metadata Overview

- Overviews

Metadata element of the following overviews are used in the global overview:

- Browsable Corpus (BC)
- Corpus Encoding Standard (CES)
- Codes for the Human Analysis of Transcripts (CHAT)
- Dublin Core (DC)
- European Language Resources Association Catalog (ELRA)
- European Science Foundation Second Language Databank (ESFSLD)
- Gesture Databank (GDB)
- International Corpus of English (ICE)
- Linguistic Data Consortium Catalog (LDC)
- Multimedia Content Description Interface (MPEG-7)
- Spoken Dutch Corpus (CGN - Corpus Gesproken Nederlands)

The following important initiatives and projects are not included in the overviews for different reasons:

- Archive of Indigenous Languages of Latin America (AILLA)

  The AILLA is a project to develop a web-based archive of linguistic materials of the indigenous languages of Latin America.

  Some info about AILLA metadata can be found here.

- Alaska Native Language Center (ANLC)

  The ANLC is recognized as the major center in the United States for the study of Eskimo and Northern Athabaskan languages. It is the center for research and documentation of the twenty Native languages of Alaska.

  There is no ANLC overview because ANLC is in the process of getting their metadata in line with Dublin Core.

- British National Corpus (BNC)

  The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written.

  The reason why there is no BNC overview is because the BNC text encoding is TEI-conformant and we rely on the Corpus Encoding Standard (CES) claim that relevant elements for corpus encoding are selected from TEI.

- Linguistic Data Archiving Project (LACITO)

  The goals of the LACITO linguistic data archiving project are the conservation and the distribution of speech data.

  The maintainers of LACITO are currently looking for a metadata standard.

- Michigan Corpus of Academic Spoken English (MICASE)

  The on-line, searchable part of a collection of transcripts of academic speech events recorded at the University of Michigan.

  The maintainers give a clear overview of meta descriptions in the form of speech events and speaker attributes.

- Text Encoding Initiative (TEI)

  The Text Encoding Initiative (TEI) is an international project to develop guidelines for the preparation and interchange of electronic texts for scholarly research, and to satisfy a broad range of uses by the language industries more generally.

TEI is not included in the overviews because we rely on the fact that the Corpus Encoding Standard selected the relevant elements from TEI for corpus encoding.

- University of Helsinki Language Corpus Server (UHLCS)

  The University of Helsinki Language Corpus Server (UHLCS) is a multilingual corpus server located at the Department of General Linguistics, the University of Helsinki. The server contains computer corpora of more than 50 languages, including samples of minority languages and extensive corpora representing different text types.

  The UHLCS is currently structured as a language hierarchy. The definition of UHLCS metadata is in progress.

These and other metadata projects that have been examined are listed in language engineering resources.

- Overview format

  Describes the format used for the overviews

- Global overview

  Gives a global overview of all metadata elements

- Language Engineering Resources

  Lists all relevant web-sites

Meta Element Overview

Version: 0.0.8 Date: 02-Oct-2000

In this overview a mapping is made between the metadata elements from different initiatives. Each column in the overview represents the elements specified by an initiative. In the first column definitions are given that agree with the definitions of the elements from different initiatives. Since the overview is made only to give an overview of the metadata used in the initiatives and not to find the exact differences between the metadata elements from the initiatives, small semantic differences between matching elements (from a row) are allowed. When the semantic difference is too large according to our knowledge, the deviating element is placed in another row and gets another definition.

The overview consists of two parts. The first part contains the elements which have a clear definition. The remaining elements in the second part of the overview are the elements without a clear definition because definitions are missing or because the given descriptions aren't specified well enough.

Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
corpus identification number			Catalog number						Description of the corpus ELRA Ref.		Identifier
corpus full title	h.title (CORPUS)	title (CORPUS)	Name		Corpus name				General Information Full name of data collection		Title
corpus short title									General Information Short name of data collection		
corpus creator	creator (CORPUS)	creator (CORPUS)							Producer Organisation, Department		Creator
corpus version	version (CORPUS)	version (CORPUS)									
corpus edition	edition (CORPUS)	release (CORPUS)									
corpus status	status (CORPUS)										

Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
corpus date created	date created (CORPUS)								Creation date		
corpus update frequency									Update frequency		
corpus date updated	date updated (CORPUS)	update (CORPUS)							Last Update		
corpus publisher	respName (CORPUS)	respName (CORPUS)							Producer Representative		Publisher
corpus publisher type	respType (CORPUS)	respType (CORPUS)							Producer Representative Position		Publisher
corpus size	wordcount, bytecount (CORPUS)	wordcount, bytecount, seccount (CORPUS)							Document Information Size		
corpus size note	extNote (CORPUS)	extNote (CORPUS)									
corpus distributor's name	distributor (CORPUS)	distributor (CORPUS)							Producer Contact Person		
corpus distributor position									Producer Contact Person Position		
corpus distributor's address	pubAddress (CORPUS)	pubAddress (CORPUS)							Producer Address, Postal Code, City, Country		

corpus distributor's telephone	telephone (CORPUS)	telephone (CORPUS)							Producer Telephone		
corpus distributor's fax	fax (CORPUS)	fax (CORPUS)							Producer Fax		
Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
corpus distributor's e-mail	eAddress.type (CORPUS)	eAddress (CORPUS)							Producer E-mail		
corpus distributor's http	eAddress. type (CORPUS)										
corpus distributor's ftp	eAddress. type (CORPUS)										
corpus identification number	idno (CORPUS)										Identifier
corpus publication date	pubDate (CORPUS)	pubDate (CORPUS)	Membership Year						Availability Date of availability		Date
corpus source title	sourceDesc h.title (CORPUS)	sourceDesc title									Source
Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
corpus source author	sourceDesc h.author (CORPUS)	sourceDesc author									
corpus source responsible name	sourceDesc respName (CORPUS)										
corpus source responsible type	sourceDesc respType (CORPUS)										
corpus source edition	sourceDesc edition (CORPUS)										

Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
corpus source identification nr.	sourceDesc idno (CORPUS)										
corpus source scope	sourceDesc biblScope (CORPUS)										
corpus source additional information	sourceDesc biblNote (CORPUS)										
corpus source publisher	sourceDesc publisher (CORPUS)	sourceDesc pubName									
corpus source publishing date	sourceDesc pubDate (CORPUS)	sourceDesc pubDate									
corpus source publishing place	sourceDesc pubPlace (CORPUS)	sourceDesc pubPlace									
corpus copyright holder		copyright (CORPUS)							Copyright Holder Organisation, Department		Rights
corpus copyright holder representative									Copyright Holder Representative		
corpus copyright holder representative position									Copyright Holder Representative Position		
corpus copyright holder contact person									Copyright Holder Contact Person		
corpus copyright holder contact person position									Copyright Holder Contact Person Position		

Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
corpus copyright holder address									Copyright Holder Address, Postal Code, City, Country		
corpus copyright holder telephone									Copyright Holder Telephone		
corpus copyright holder fax									Copyright Holder Fax		
corpus copyright holder e-mail									Copyright Holder E-mail		
corpus description	projectDesc (CORPUS)	projectDesc (CORPUS)			Corpus description, fdescription, infofile				General Information Source, Additional Information Documentation, On-line documentation		Description
corpus type			Corpus Type						General Information Type of resource		Type
corpus ISBN			ISBN								Identifier
information about the methods for text sampling in the corpus	samplingDecl								Document Information Details of the source, Domain, Time Span		
corpus standard conformance level	conformance.level (CORPUS)	conformance.level									
corpus language used	profileDesc language (CORPUS)	profileDesc language	Language		Content Language				Document Information Language(s)		Language
corpus character set used	profileDesc wsdUsage (CORPUS)	profileDesc wsdUsage							Technical Information Character set		

Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
corpus distribution media									Technical Information distribution media		
corpus related tools									Additional Information Related Tools		
corpus application purposes									Additional Information Application purposes		
corpus keyword	profileDesc keyTerm (CORPUS)				Content Keywords						
corpus revision description	revisionDesc change (CORPUS)	revDesc resp									
corpus revision date	revisionDesc changeDate (CORPUS)	revDesc date									
corpus revision name of responsible	revisionDesc respName (CORPUS)	revDesc respName									
corpus price									Availability Price for research use, Price for commercial use	InternationalPriceType, Currency, Value	
document title	h.title (DOC)	title (DOC)								Creation TitleText	Title
document description					Content description, infofile					Creation Abstract	Description
document type				Text categories						Classification Subject	Type
document genre										Classification Genre	
document name of the creator	creator (DOC)	creator (DOC)						Researcher Name		Creation CreatorType	Creator
document role of the creator										Creation Role	

document version number	version (DOC)	version (DOC)									
document revision number	editionStmt version (DOC)										
document status description	status (DOC)										
document date of creation	date.created (DOC)										Date
Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
document date of last update	date.updated (DOC)	update (DOC)									Date
document name of publisher	respName (DOC)	respName (DOC)									Publisher
document type of publisher	respType (DOC)	respType (DOC)									
document size	wordcount, bytecount (DOC)	wordcount, seccount, bytecount (DOC)									
document size information	extNote (DOC)	extNote (DOC)									
document distributor's name	distributor (DOC)	distributor (DOC)								UsageRecord Distributor	
document distributor's address	pubAddress (DOC)										
document distributor's telephone	telephone (DOC)										
Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
document distributor's fax	fax (DOC)										
document distributor's e-mail	eAddress.type (DOC)										

Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
document distributor's http	eAddress. type (DOC)										
document distributor's ftp	eAddress. type (DOC)										
document identification number	idno (DOC)										Identifier
document date of publication	pubDate (DOC)										Date
document bibliographic source title	sourceDesc h.title (DOC)	sourceDesc title									Source
document bibliographic source author	sourceDesc h.author (DOC)	sourceDesc author									
document bibliographic source name of responsible	sourceDesc respName (DOC)										
document bibliographic source type of responsible	sourceDesc respType (DOC)										
document bibliographic source edition	sourceDesc edition (DOC)										
document bibliographic source identification number	sourceDesc idno (DOC)										
document bibliographic source scope	sourceDesc biblScope (DOC)										
document bibliographic source additional information	sourceDesc biblNote (DOC)										
document bibliographic source name of publisher	sourceDesc publisher (DOC)	sourceDesc pubName									

Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
document bibliographic source publishing date	sourceDesc pubDate (DOC)	sourceDesc pubDate									
document bibliographic source publishing place	sourceDesc pubPlace (DOC)	sourceDesc pubPlace									
document source date of recording		sourceDesc rec date					Date of Encounter	Recording Date			
document source time of recording		sourceDesc rec time									
document source description		sourceDesc source									
document source producer		sourceDesc producent									

Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
project name			Research Project					Project Name			
project controller identification							Interviewer				
project controller name					Project Controller Name						
project controller description	projectDesc (DOC)				Project Controller description, infofile						
document standard conformance level	conformance.level (DOC)	conformance.level									

Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
document origination type			Data sources	TV or Radio							

Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
document origination description	profileDesc creation (DOC)										
document origination access rights					Access					Rights RightsType	Rights
document origination access date					Access date						Date
document origination media identification					Tape ID		Cassette / Recording	Video Identifier			
document origination media position					Tape Position			Start Position, End Position			
document origination media description					Tape Description						
document origination media format		recCondition recMedium type			Tape Format					MediaFormat Medium	Format
document origination compression format										MediaCoding CompressionFormat	
document origination media link										MediaInstance InstanceLocator	
document origination recording device										CreationMaterial DeviceInstrument	
document origination recording device settings										CreationMaterial DeviceSettings	
document origination recording location										Classification Country	
document origination size								Filesize		MediaFormat FileSize	
Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC

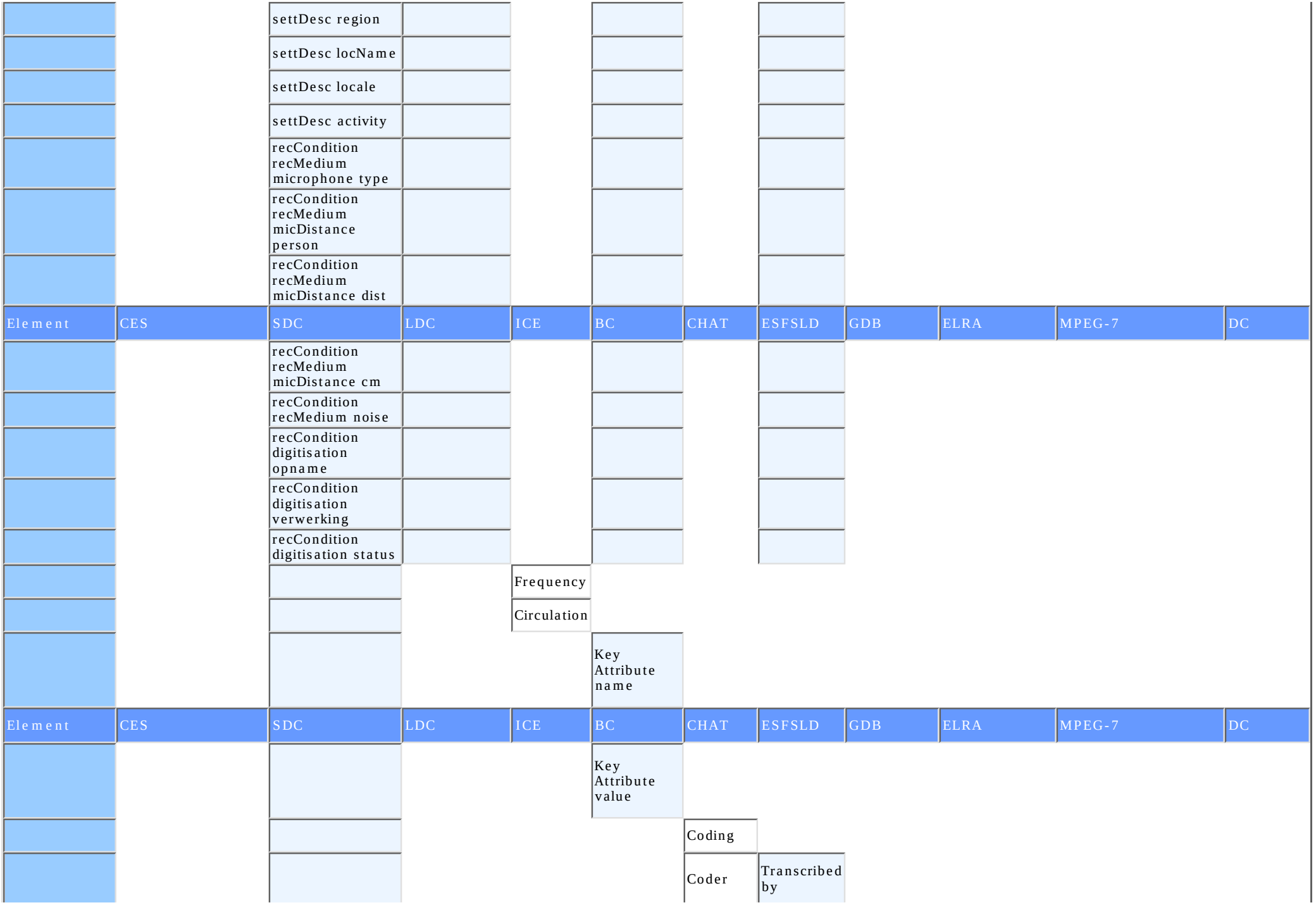
Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
document language used	profileDesc language (DOC)	profileDesc language			Content Language		Informant Source Language			Classification Language	Language
document character set used	profileDesc wsdUsage (DOC)	profileDesc wsdUsage									
document keyword	profileDesc keyTerm (DOC)	-			Content Keywords		Keywords				
document revision description	revisionDesc change (DOC)	revDesc resp					History				
document revision date	revisionDesc changeDate (DOC)	revDesc date									
document revision name of responsible	revisionDesc respName (DOC)	revDesc respName					Revised by / Checked by				
corpus recommended application			Recommended Application								
participant's identification		particDesc id			Participants Code	Speaker's ID	Informant / Subject	Participant's Abbreviation			Identifier
participant's name					Participants Name, Fullname	Speaker's Name		Participant's Name			
participant's date of birth		particDesc year			Participants Born	Speaker's Date of Birth	Subject's Date of Birth				Date
participant's age		particDesc age		Speakers Age	Participants Age						
participant's gender		particDesc sex		Speakers Gender	Participants Sex	Speaker's Sex	Subject's Sex				
participant's role		particDesc role		Speakers role	Participants Role	Speaker's Role					
participant's education		particDesc education				Speaker's Education					

participant's socio-economic status (SES)						Speaker's SES					
participant's religion							Subject's Religion				
participant's regional scope				Scope							
number of participants				Speakers per text							

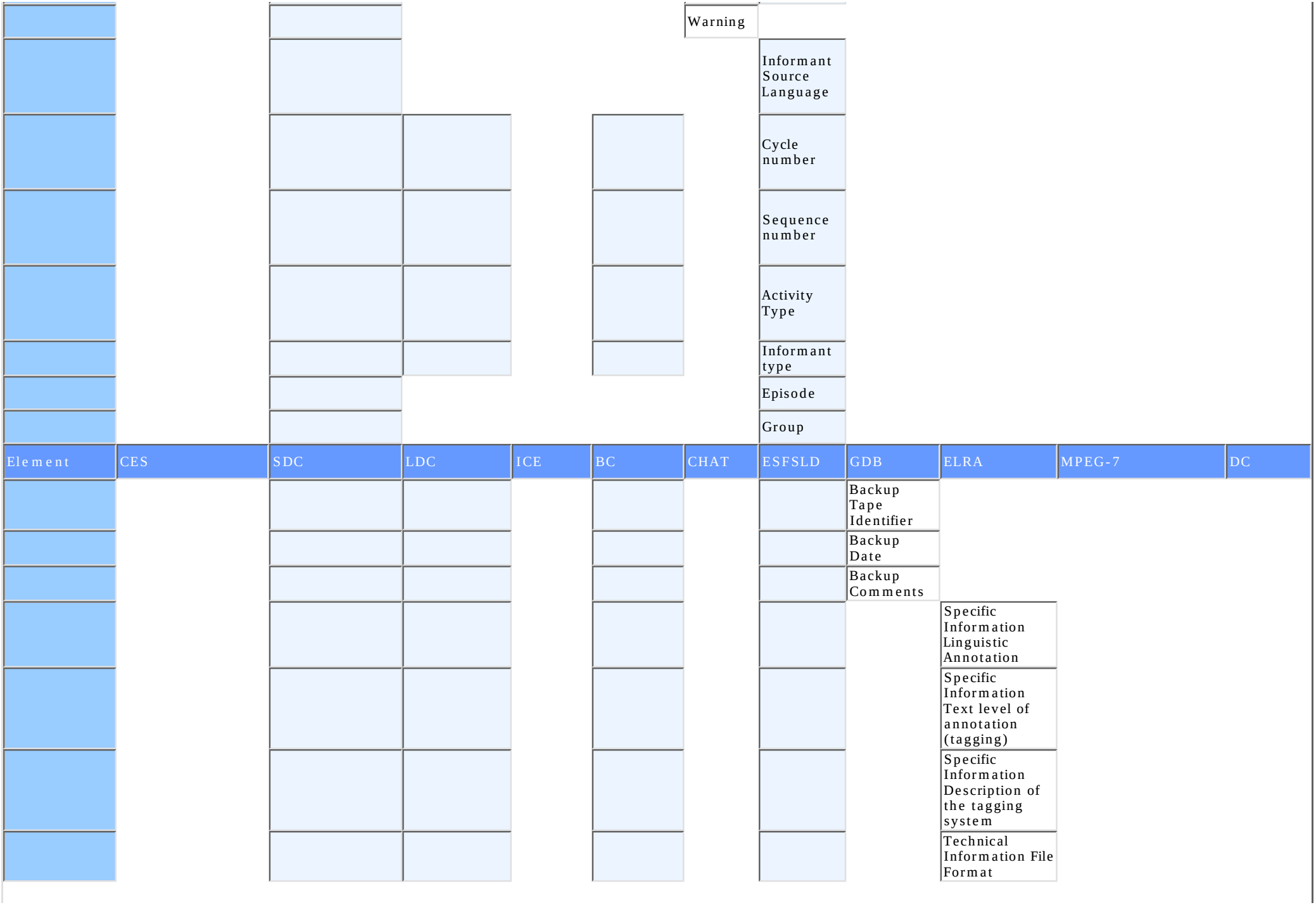
Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
	sourceDesc imprint	sourceDesc imprint									
	availability	availability									
	availability.region	availability.region									Coverage
	availability.status	availability.status									
	profileDesc creation (CORPUS)										
	profileDesc catRef	profileDesc catRef									
	profileDesc translation (DOC/CORPUS)	profileDesc translation									
	profileDesc translator (DOC/CORPUS)	profileDesc translator									

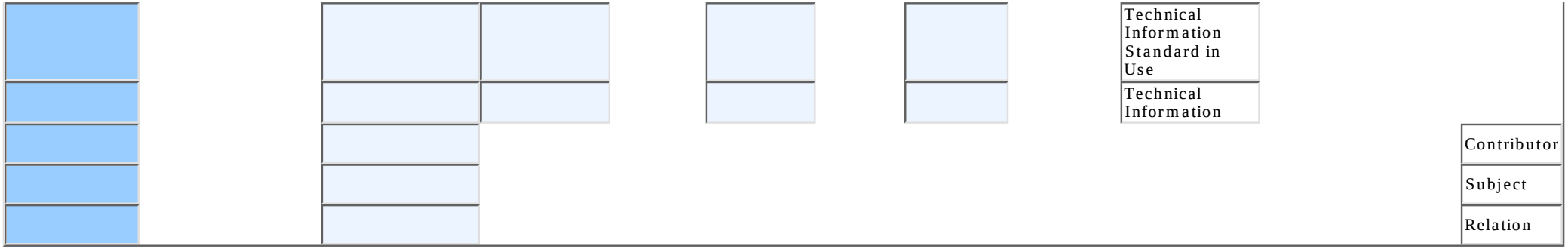
Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
	profileDesc annotation.type (DOC/CORPUS)	profileDesc annotation.type									
	profileDesc annotation ann.loc (DOC/CORPUS)	profileDesc annotation ann.loc									

	profileDesc annotation trans.loc (DOC/CORPUS)	profileDesc annotation trans.loc									
					Participants Relation						
Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
	transduction (DOC/CORPUS)	transduction									
	correction (DOC/CORPUS)										
	quotation (DOC/CORPUS)										
	hyphenation (DOC/CORPUS)										
	segmentation (DOC/CORPUS)	segmentation									
	normalization (DOC/CORPUS)										
	tagUsage.gi (DOC/CORPUS)										
Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
	tagUsage.occurs (DOC/CORPUS)										
	tagUsage.wsd (DOC/CORPUS)										
	refsDecl (DOC/CORPUS)	refDecl ?									
	category (DOC/CORPUS)	category									
	catDesc (DOC/CORPUS)	catDesc								Classification ClassificationType	
		particDesc interaction type									
		particDesc interaction active									
		particDesc interaction passive									
Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
		particDesc relation active									
		particDesc relation desc									
		particDesc relation mutual									



		settDesc region									
		settDesc locName									
		settDesc locale									
		settDesc activity									
		recCondition recMedium microphone type									
		recCondition recMedium micDistance person									
		recCondition recMedium micDistance dist									
Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
		recCondition recMedium micDistance cm									
		recCondition recMedium noise									
		recCondition digitisation opname									
		recCondition digitisation verwerking									
		recCondition digitisation status									
				Frequency							
				Circulation							
					Key Attribute name						
Element	CES	SDC	LDC	ICE	BC	CHAT	ESFSLD	GDB	ELRA	MPEG-7	DC
					Key Attribute value						
						Coding					
						Coder	Transcribed by				


Warning
Informant Source Language
Cycle number
Sequence number
Activity Type
Informant type
Episode
Group
Element
CES
SDC
LDC
ICE
BC
CHAT
ESFSLD
GDB
ELRA
MPEG-7
DC
Backup Tape Identifier
Backup Date
Backup Comments
Specific Information Linguistic Annotation
Specific Information Text level of annotation (tagging)
Specific Information Description of the tagging system
Technical Information File Format


Technical Information Standard in Use
Technical Information
Contributor
Subject
Relation

Browsable Corpus (BC)

Introduction	References	Corpus Structure	Corpus Information
Document Information	Header Information	Metadata Overview	

Last update: 30-Aug-2000

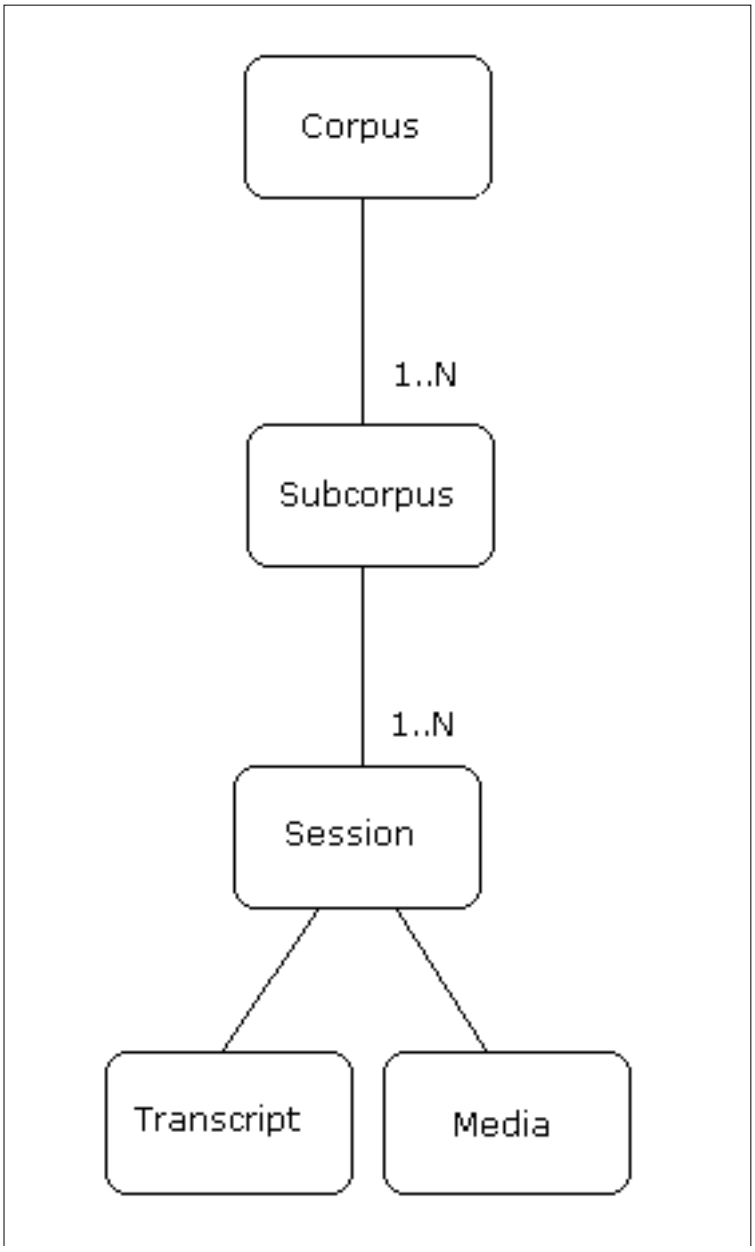
Introduction

The Browsable Corpus concept was introduced at the Max Planck Institute for Psycholinguistics (MPI) to make resource discovery easier by defining meta-descriptions for language resources. The structure of linked meta-descriptions can be browsed and searched.

References

A Browsable Corpus: accessing linguistic resources the easy way (Broeder, Brugman, Russel & Wittenburg).

Corpus Structure



Metadata Overview

Corpus	Groups corpora and/or sessions together			
	Name	Name of the grouping		
	Level	Not used anymore		
	Description	Description of the grouping		
0..N	FDescription	Provides for (legacy) HTML information files		
0..N	Infofile	Reference to a file/URL providing relevant information (usually legacy metadata) about the current grouping		
	Project_Controller	Responsible for the corpus data		
		Name	Name of the entity responsible for the corpus data	
		Description	Description of the responsible entity	
		Infofile	Reference to a file/URL providing information about the responsible entity	
	Content	Describes important aspects about the content of the corpus		
		Keywords	Keywords about the content	
		Languages	Groups the languages used in the conversation	
		1..N	Language	The language used in the conversation
			Description	Describes specifics about the language used in the conversation
			Infofile	Reference to a file/URL providing information about the language used in the conversation
		Description	Description of the content of the corpus	
	0..N	Infofile	Reference to a file/URL providing information about the contents of the corpus	
	Participants	Groups together information about the persons identified in the corpus		
	1..N	Person	Gives information about an identified person	
			Name	Name of the person
			Fullname	Person's full name
			Sex	Person's gender
			Role	Person's role in the conversation
			Code	Person's unique identifier
			Age	Person's age
			Born	Person's date of birth
			Relation	?

			Keys	Contains a set of attribute-value pairs		
			0..N	Key	Attribute-value pair	
					Name	Contains the attribute label name
					Value	Contains the attribute's value
		Description	Description of the participants			
	0..N	Infofile	Reference to a file/URL providing information about the participants			
	Keys	Contains a set of attribute-value pairs				
		0..N	Key	Attribute-value pair		
			Name	Contains the attribute label name		
			Value	Contains the attribute's value		
Session	?					
	Name	**** see corpus ****				
	Level	**** see corpus ****				
	Date	**** see corpus ****				
	Description	**** see corpus ****				
0..N	FDescription	**** see corpus ****				
	Access	Contains juridical access rights to the recording				
		Date	Contains the date of when the rights were established			
0..N	Infofile	**** see corpus ****				
	Project_Controller	**** see corpus ****				
	Content	**** see corpus ****				
	Participants	**** see corpus ****				
0..N	Tape	Contains a possible reference to the original media tape				
		Description	Description of the media tape			
		ID	Gives a unique identifier to locate the tape in the archive			
		Position	Indicates the position on the tape where the session resides			
		Format	Describes the media format on which the recording is made			
	Keys	**** see corpus ****				
	Files	Groups together all the files of the session				
	0..N	Transcription-File	Groups information about a transcription file			
			Remark	Description of the transcription file		

			Src	Points to the location of the transcription file
			Format	Indicates how the file should be interpreted
	0..N	Label_File	Gives a sequence of markers which relate to the media file fragment	
			Remark	Description of the label file
			Src	Points to the location of the label file
			Format	Indicated how the file should be interpreted
			Start	Gives information about the start position of media file fragment
			Duration	Gives information about the duration of the media file fragment
	0..N	Media-File	Groups information about a media file	
			Remark	Description of the media file
			Src	Points to the location of the media file
			Format	Gives the format of the media file
			Start	Indicates when the recording was started
			Duration	Gives information about the duration of the recording
			Audio-Quality	Gives information about the audio quality of the recording
			Video-Quality	Gives information about the video quality of the recording
	0..N	Infofile	**** see corpus ****	

Corpus Encoding Standard (CES)

Last update: 30-Aug-2000

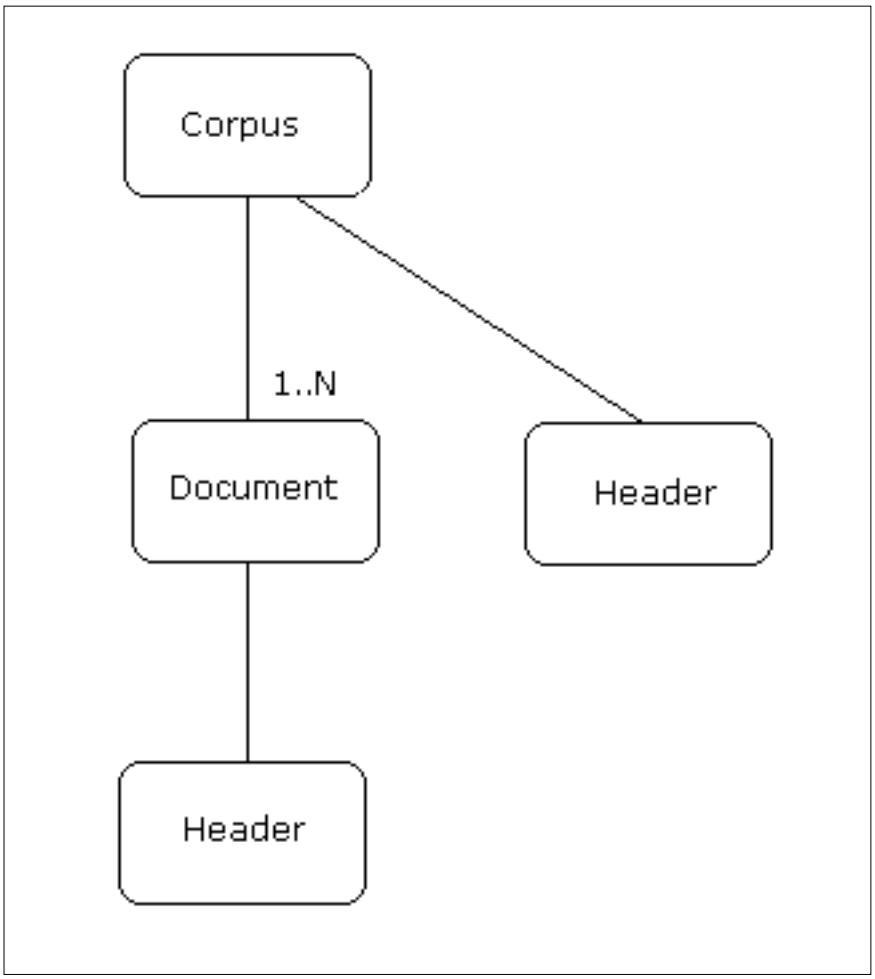
Introduction

The Corpus Encoding Standard (CES) is an encoding standard for corpus-based work for use in the language engineering community. The CES is an application of SGML and conformant to the TEI guidelines.

References

Information about the CES was taken from the CES document version 1.4 (Nancy Ide, 1996).

Corpus Structure



Corpus Information

A CES encoded corpus contains a single corpus header (cesHeader) and one or more documents (cesDOC). Each document contains a single text header and a text. Additionally, the cesCorpus element can be recursively nested, and sequences of this element can appear at any nested level, in order to identify sub-corpora.

Document Information

A document, defined by cesDoc, contains a header (cesHeader) followed by either a <body> element or a <group> element.

Header Information

The header (cesHeader) provides information about the electronic text that has been encoded, including not only its title, author etc. but also information about its encoding. The elements in the header are:

- fileDesc: Bibliographic description of the corpus or document. This information is required.
- encodingDesc: Documents the relationship between an electronic text and the source or sources from which it was derived
- profileDesc: Provides further information about various aspects of a text, specifically the language used, the situation and date of its production, the participants and their setting, and a descriptive classification for it
- revisionDesc: Summarizes the revision history for a file

Metadata Overview

type *	The kind of document to which the header is attached. CORPUS when the header is attached to the corpus and TEXT when attached to a single text.		
creator *	The agency responsible for creating the header.		
version *	The version and revision of the CES header.elt used to encode this header. This number is found near the top of the header.elt itself		
status *	The revision status of the header. NEW when it is the first version of the header and UPDATE when the header has been updated.		
date.created *	The date on which the header content was created.		
date.updated *	The date on which the header content was last updated.		
fileDesc	Contains a full bibliographic description of the corpus itself or of a text within it. The elements contained are: titleStmt, editionStmt, extent, publicationStmt and SourceDesc. The elements titleStmt, publicationStmt and sourceDesc are required.		
	titleStmt	Groups information concerning the title of the corpus or the individual text and its constituent texts.	
		h.title	The title of the electronic file, including alternative titles or subtitles.
		respStmt	supplies information about any person or institution responsible for the intellectual content of a text, edition, or electronic transcription.
			respType: contains a phrase describing the nature of person's or institution's intellectual responsibility
			respName: the publisher of the corpus or text expressed as the proper name of a person, place or institution.
	editionStmt	Contains any additional information relating to a particular version of a text.	
		version	
	extent	provides the size of the electronic text as stored on some carrier medium.	
		wordCount	contains the count of words in the text

	byteCount	contains the count of bytes in the file containing the text together with its markup.	
		units *	Gives the unit in which the bytecount is measured (BYTES : bytes, KB : kilobytes, MB : megabytes, GB: gigabytes)
	extNote	A descriptive note supplying additional information of any kind relating to an extent information provided within a corpus or text header.	
publicationStmt	Groups information concerning the publication or distribution of the corpus and its constituent texts.		
	distributor	Gives the name of the person or institution who distributes the text or corpus	
	pubAddress	Contains the postal address of the distributor	
	telephone	Gives the telephone number of the person or institution who distributes the text or corpus, in format conformant to ITU-T/CCITT Recommendation E.123	
	fax	Gives the fax number of the person or institution who distributes the text or corpus, in format conformant to ITU-T/CCITT Recommendation E.123	
	eAddress	Gives an electronic address of the person or institution who distributes the text or corpus. Note that more than one occurence of this tag can appear, so that multiple addresses (possibly of different types) can be included	
		type *	Gives the type of the electronic address (email address, web site, ftp site, etc.)
	availability	Supplies information about the availability of a text, for example, any restrictions on its use or distribution, its copyright status, etc	
		region *	specifies the territories within which rights in the electronic text apply
		status *	supplies a code identifying the current availability of the text
	idno	Supplies a number (e.g., ISBN) used to identify a bibliographics item	
	pubDate	The publication date expressed in any format	
		value *	Specifies standard value for this date in ISO 8601 (Representation of dates and times) format
sourceDesc	Supplies a bibliographic description of the copy text(s) from which an electronic text was derived or generated		
1..N	biblStruct	Contains a structured bibliographic citation, in which only bibliographic sub-elements appear and in a specified order	
		analytic	Contains bibliographic elements describing an item (e.g. an article or poem) published within a monograph, journal, or periodical and not as an independent publication
		monogr	Contains bibliographic elements descibing an item (e.g. a book or journal) published as an independent item (i.e. as a separate physical object).
			h.title the title of a work

h.author	in a bibliographic reference, contains the name of an author (personal or corporate) of a work; names should be given in a canonical form, with surnames preceding forenames	
respStmt	supplies information about any person or institution responsible for the intellectual content of a text, edition, or electronic transcription	
edition	Provides bibliographic details for an edition of some text	
imprint	groups information relating to the publication or distribution of a bibliographic item	
idno	Supplies a standard (e.g., ISBN) number used to identify a bibliographic item	
	type *	A name of abbreviation (e.g., ISBN) identifying what type of identifying number is given. Unless provided explicitly the default value is: ISBN
biblScope	Defines the scope of a bibliographic reference, for example as a list of page numbers, or a named subdivision or a larger work.	
	type *	Identifies the type of information conveyed by the element (PP : page number or page range, VOL : volume number, ISSUE : issue number)
biblNote	A descriptive not supplying additional information of any kind relating to a bibliographic item described wihtin a corpus or text header	
publisher	Proper name of a person, place or institution	

			type *	categorises the name (PERSON : name of person, PLACE : name of a place, ORG : name of an organization article in a periodical)
		pubDate	A calendar date in any format	
			value *	Specifies standard value for this date in ISO 8601 format
		pubPlace	Place of publication for a book, article, etc	
encodingDesc	Documents the relationship between an electronic text and the source or sources from which it was derived			
	projectDesc	Describes in detail the purpose for which an electronic file was encoded		
	samplingDecl	Contains a prose description of the rationale and the methods used in sampling text in the creation of the corpus		
	editorialDecl	Provides details of editorial principles and practices applied during the encoding of a text		
		conformance	Provides the CES level of conformance for the text or corpus	
			level *	Gives the level of CES conformance (legal values are 1, 2 or 3)
		transduction	Describes the principles according to which the text has been transduced, either in transcribing it from audio tape to written form, or in converting from an electronic original	
		correction	Specifies a set of correction practices applied in creating one or more components of the corpus	
		quotation	Specifies editorial practice adopted with respect to qoutation marks in the original	
			marks *	Indicates whether or not quotation marks are retained as tag content in the text (NONE : no quotation marks retained, SOME: some quotation marks retained, ALL : all quotation marks retained)
			form *	Specifies how quotation marks are indicated within the text (STD : use of quotation marks has been standardized; open and close quote marks are distinct, NONSTD : open and close quote marks are represented indiscriminately by the ????? , UNKNOWN : use of quotation marks unknown)
		hyphenation	Summarizes the way in which end-of-line hyphenation in a source text has been treated in an encoded version of it	
		segmentation	Describes the principles according to which the text has been segmented, for example into sentences, tone-units, graphemic strata, etc	
		normalization	Specifies a set of normalization practices applied in creating one or more components of the corpus	
			method *	Indicates whether normalization made without notation or made by including editorial tags (TAGS : normalization indicated with tags, SILENT : normalization made silently)

tagsDecl	Provides detailed information about the tagging applied to an SGML document			
1..N	tagUsage	Supplies information about the usage of a specific element within the corpus or text with which this header is associated		
			gi *	The name (generic identifier) of the element indicated by the tag
			occurs *	Specifies the number of occurrences of this element within the text
			wsd *	Can be used on a <tagUsage> element to indicate that for every appearance of the described element in the text, the content defaults to the specified character set
refsDecl	Specifies how canonical references are constructed for this text			
classDecl	Contains a series of <category> elements, defining the classification codes used for texts within the corpus			
1..N	taxonomy	Defines a typology used to classify texts		
	1..N	category	Contains an individual descriptive category or feature-value pair	
			catDesc	Describes a category within the text typology, in the form of a brief prose description
profileDesc	Provides further information about various aspects of a text, specifically the language used, the situation and date of its production, the participants and their setting, and a descriptive classification for it			
creation	Contains information about the origination of a text			
langUsage	Groups information describing the languages, sublanguages, registers, dialects etc. represented within a text			
1..N	language	Characterizes a language, sublanguage, register, dialect, etc., used within a single text		
			iso639 *	Gives the standard language code from ISO 639 in one of the following forms: a two-letter code from ISO 639, a three-letter code from ISO 639-2 or one of the above extended by a country code from ISO 3166
			type *	Indicates the type of language, e.g., sublanguage, dialect, etc
wsdUsage	Groups information describing the character set(s) used within a text			
1..N	writingSystem	Characterizes a character set used within a single text		
textClass	Groups information which describes the nature or topic of a text in terms of a standard classification scheme, thesaurus, etc			
	catRef	Specifies one or more defined categories within some taxonomy or text typology		
			target *	Identifies the text category or categoeries, by means of an IDREF pointing to one or more <category> elements defined in the corpus header
			scheme *	identifies the classification scheme
	h.keywords	Contains a list of keywords or phrases identifying the topic or nature of a text, each of which is tagged as a term. A standard list will be provided by EAGLES/PAROLE		
	1..N	keyTerm	Contains a technical term or phrase, particularly in a list of descriptive keywords	
translations	Groups information about existing translations of the text			

	1..N	translation	Gives information about a translation of the text. The global lang attribute and the wsd attribute are required on this tag	
			trans.loc *	Provides information (path/file name, URL, etc.) about the location of the translation
		translator	Gives the name of the translator	
	annotations	Groups information about existing annotation files associated with the text		
	1..N	annotation	Gives information about an annotation file associated with the text	
			type *	Indicates the type of annotation (SEGMENT : annotation file contains segmentation into sentences and words, GRAM : annotation file contains morpho-syntactic category information for the words in the text, ALIGN : annotation file contains alignment links to a parallel translation
			ann.loc *	Provides information (path/file name, URL, etc.) about the location of the annotation file
			trans.loc *	For annotation files containing alignment information, provides information (path/file name, URL, etc.) about the location of the file containing the aligned text
revisionDesc	Summarizes the revision history for a file			
1..N	change	Summarizes a particular change or correction made to a particular version of an electronic text which is shared between several researchers		
		changeDate	Gives the date of the change	
			value *	Specifies standard value for this date in ISO 8601 format
		respName	Specifies the person responsible for the change	
		h.item	Specifies the nature of the change(s). One or more occurrences of this element may appear within each <change> element	

* *attribute*

Codes for the Human Analysis of Transcripts (CHAT)

Last update: 30-Aug-2000

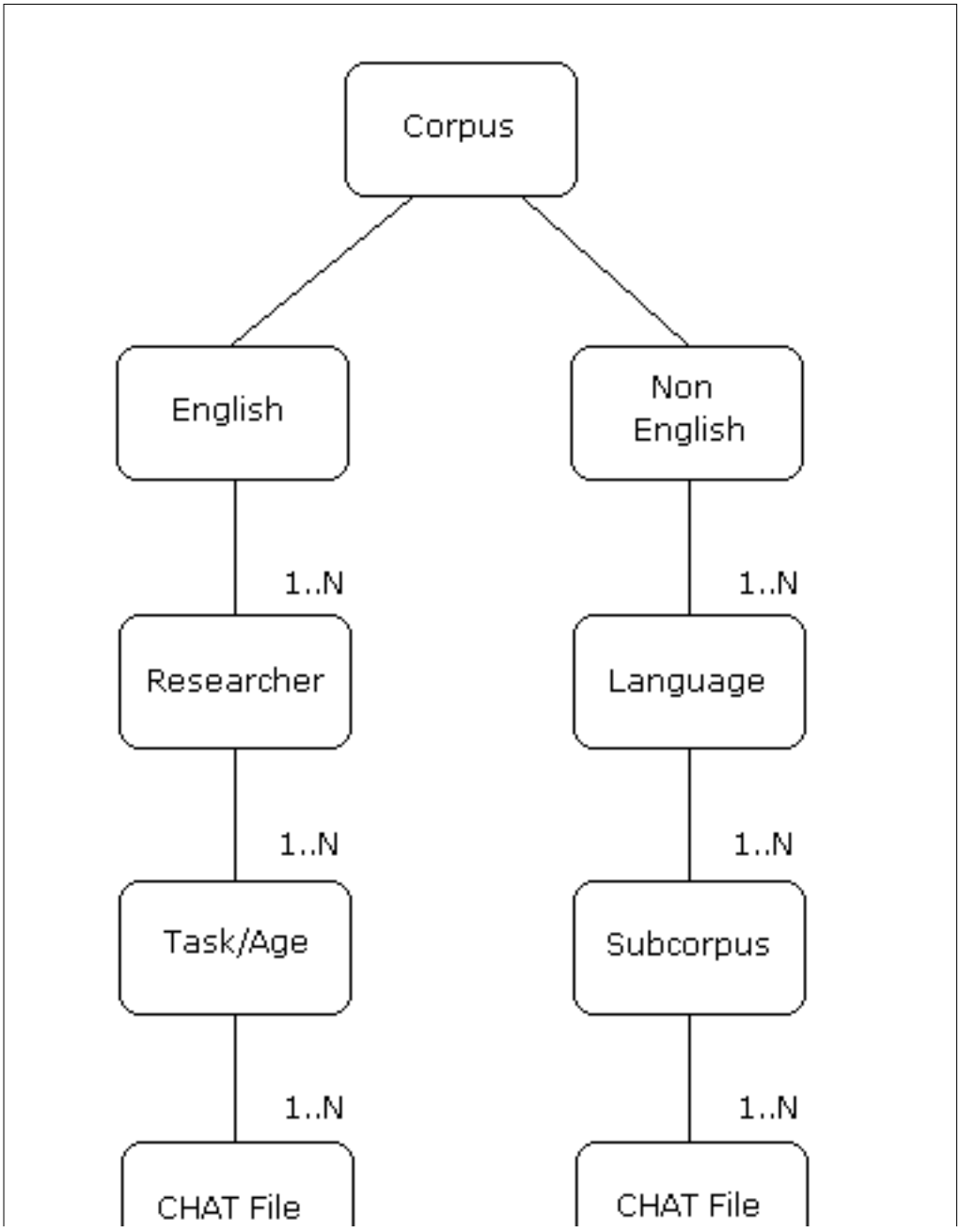
Introduction

CHAT is the format used for the CHILDES (Child Language Data Exchange System) project.

References

MacWhinney, Brian. 1991. The Childes Project: Tools for Analyzing Talk

Corpus Structure



Corpus Information

The corpus header is called the 'Documentation File' in CHAT. It is stored in a text file (00readme.doc) in the corpus directory. The documentation file contains descriptions about the corpus. Some metadata elements which are extracted from these human readable descriptions are listed under 'Corpus Header' in the metadata overview.

Document Information

Each document equals one file in the corpus directory.

Header Information

There are three types of document headers in CHAT:

- Obligatory headers
- Constant headers
- Changeable headers

Metadata Overview

Corpus Header	A basic set of facts about the corpus		
	Acknowledgements	A statement that asks the user to cite some particular reference when using the corpus	
		Reference Name	The name of the person cited
		Reference Year	The year of the cited reference
	Restrictions	A description of the restrictions on the use of the corpus data	
	Warnings	A description of the limitations on the use of the corpus data	
	Pseudonyms	?	
	History	Gives detailed information about the history of the project	
		Funding	Description of how the funding was obtained
		Goals	Description of the goals of the project
		Data collection	Description of how the data was collected
		Sampling procedure	Description of the sampling procedure
		Transcription procedure	Description of the transcription procedure
		Transcription ignored	Description of what was ignored in the transcription
		Transcribers training	Description of the transcribers training
		Reliability	Description of reliability of the data
		Coding	Description about coding and used codes ????
		Computerized	Description of how the material was computerized ????
	Codes	Description of project-specific codes	

	Biographical data	Gives biographical information about the informant	
		Informant's Age	?
		Informant's Gender	?
		Informant's Siblings	?
		Informant's Schooling	?
		Informant's Social Class	?
		Informant's Occupation	?
		Informant's Previous residences	?
		Informant's religion	?
		Informant's interest	?
		Informant's friends	?
	Table of contents	Gives a brief index to the contents of the corpora ?	
	Situational description	Gives general situational descriptions ?	
Obligatory header	This header must be included to for use with CLAN programs		
	Participants	Lists all the 'actors' within the file	
	1..N	Speaker's ID	The participants are represented by a unique three-letter ID. Mostly the first three letters from the speaker's name are used
		Speaker's Name	The speaker's first name
		Speaker's Role	The speaker's relationship to the children under study. Standard roles: Target_Child, Mother, Father, Brother, Sister, Teacher, Playmate and Investigator
Constant header	Contains information that is constant throughout the file. The information is unlikely to change during the course of the recording session		
	Age	Specifies the speaker's age in years, months and days.	
	1..N	Speaker's ID	The unique speaker's ID which refers to the name and role of a participant
		Speaker's Age	Age in years, months and days
	Birth	Gives the date of birth of the speaker	
	1..N	Speaker's ID	The unique speaker's ID which refers to the name and role of a participant
		Speaker's Date of birth	Date of birth
	Coding	Indicates the date of the current version of CHAT. Used for updating files and new coding conventions	
	Coder	Identifies the people who transcribed and coded the file	
	Education	Specifies the speaker's highest grade in school	
	1..N	Speaker's ID	The unique speaker's ID which refers to the name and role of a participant
		Speaker's Education	Identifies the speaker's education or years of college
	Filename	Gives the name of the computer file	
	ID	Used by the program "STATFREQ" to assign a unique code to each child	
	1..N	Speaker's ID	The unique speaker's ID which refers to the name and role of a participant
		Unique code	A unique code to identify the speaker throughout a corpus
	SES	Describes the socioeconomic status of the child's family	
	1..N	Speaker's ID	The unique speaker's ID which refers to the name and role of a participant

		Speaker's SES	The speaker's socioeconomic status. The following adjectives are recommended: welfare, lower, working, lower-middle, middle, upper-middle, upper
	Sex	Indicates the speaker's gender	
	1..N	Speaker's ID	The unique speaker's ID which refers to the name and role of a participant
		Speaker's Sex	Gender of the speaker (male or female)
	Warning	Describes user warnings about certain defects or peculiarities in the collection	
Changeable header	Contain information that can change within the file		
	Activities	Describes the activities involved in a situation	
	Bgd	Describes backgrounding material (????)	
	Comment	Used for all-purpose comments	
	Date	Indicates the date of interaction	
	Language	Specifies the language used for the material that follows	
	Location	Indicates the city, state and country in which the interaction took place	
	New Episode	Indicates the end of one episode and the beginning of another	
	Room Layout	A description of the room and its contents	
	Situation	Describes the general setting of the interaction	
	Stim	Indicates a particular stimuli used in an elicited production task	
	Tape Location	Indicates the specific tape from which the transcription was made	
		Tape ID	Gives the tape identifier
		Tape Side	Gives the side of the tape (a or b)
		Tape footage	Gives the tape footage
	Time Duration	Indicates the time at which the audiotaping began and the time that passed during the course of the taping	
		Time Start	Gives the time at which the recording began
		Time End	Gives the time at which the recording ended
	Time Start	Used to "restart" the clock	

The Dublin Core (DC)

Introduction	References	Corpus Structure	Corpus Information
Document Information	Header Information	Metadata Overview	

Last update: 09-Aug-2000

Introduction

The Dublin Core is a metadata element set intended to facilitate discovery of electronic resources. Originally conceived for author-generated description of Web resources, it has attracted the attention of formal resource description communities such as museums, libraries, government agencies, and commercial organizations.

References

Information about the Dublin Core comes from the web-page (The Dublin Core MetaData Initiative, http://purl.oclc.org/dc/).

Metadata Overview

Title	A name given to the resource
Creator	An entity primarily responsible for making the content of the resource
Subject	The topic of the content of the resource
Description	An account of the content of the resource
Publisher	An entity responsible for making the resource available
Contributor	An entity responsible for making contributions to the content of the resource
Date	A date associated with an event in the life cycle of the resource
Type	The nature or genre of the content of the resource
Format	The physical or digital manifestation of the resource
Identifier	An unambiguous reference to the resource within a given context
Source	A Reference to a resource from which the present resource is derived
Language	A language of the intellectual content of the resource
Relation	A reference to a related resource
Coverage	The extent or scope of the content of the resource
Rights	Information about rights held in and over the resource

European Language Resources Association Catalog (ELRA)

Introduction	References	Corpus Structure	Corpus Information
Document Information	Header Information	Metadata Overview	

Last update: 25-Sep-2000

Introduction

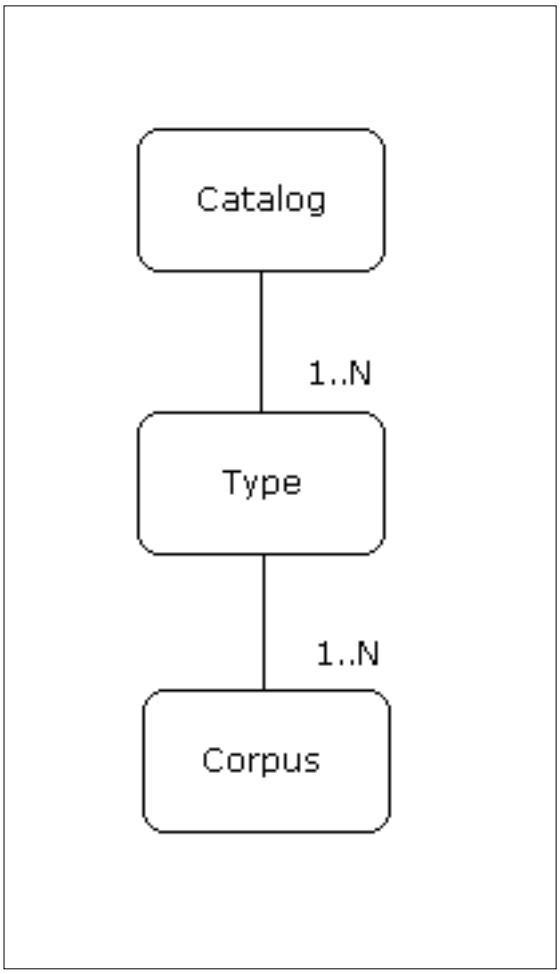
ELRA (European Language Resources Association) is an organization to promote the creation, verification, and distribution of language resources in Europe. The catalog includes a wide range of corpora including speech corpora, written corpora and terminology corpora.

References

Information for this overview comes from the "written corpus" description form found at the European Language Resources Association Catalog

Catalog Structure

The catalog is an access structure on top of corpora where the metadata is about the corpora in the catalog. Corpora are divided into categories according to the type of data they contain (speech, written, terminology). Each category contains a set of corpora.



Meta Data Overview

Producer / Provider		
	Organisation	
	Department	
	Representative	
	Representative Position	
	Contact Person	
	Contact Person Position	
	Address	
	Postal Code	
	City	
	Country	
	Telephone	
	Fax	
	E-mail	
Copyright Holder		
	Organisation	
	Department	
	Representative	
	Representative Position	
	Contact Person	
	Contact Person Position	
	Address	
	Postal Code	
	City	
	Country	
	Telephone	
	Fax	
	E-mail	
General Information		
	Full name of data collection	
	Short name of data collection	
	Type of resource	According to the description form the following types are possible: Monolingual, Multilingual, Parallel and Other.

	Source	
	Creation date	
	Update frequency	
	Last update	
Document information (?)		
	Language(s)	
	Details of the source	
	Domain(s)	
	Size (in words, sentences, etc.)	
	Time span	
Description of the corpus		
	ELRA Ref.	
Specific Information		
	Linguistic Annotation	Type of annotation? (Phonemic, Othographic, Morphological, Syntactical, Semantic, Other)
	Text level of annotation (tagging)	
	Description of the tagging system	
Technical Information		
	File format	Text, Word for PC, Word for Mac, Other
	Standard in use	ISO, SGML, TEI, Other
	Character set	ISO 8859-1, 7-bit ASCII, 8-bit ASCII, UNICODE, Other
	Distribution media	CD-ROM, Floppy disk, Cartridge, Other
Additional Information		
	Documentation	
	On-line documentation (www, ftp)	
	Related tools	
	Application purposes	
Availability		
	Date of availability	
	Price for research use	
	Price for commercial use	

European Science Foundation Second Language Databank (ESFSLD)

Introduction	References	Corpus Structure	Corpus Information
Document Information	Header Information	Metadata Overview	

Last update: 18-Sep-2000

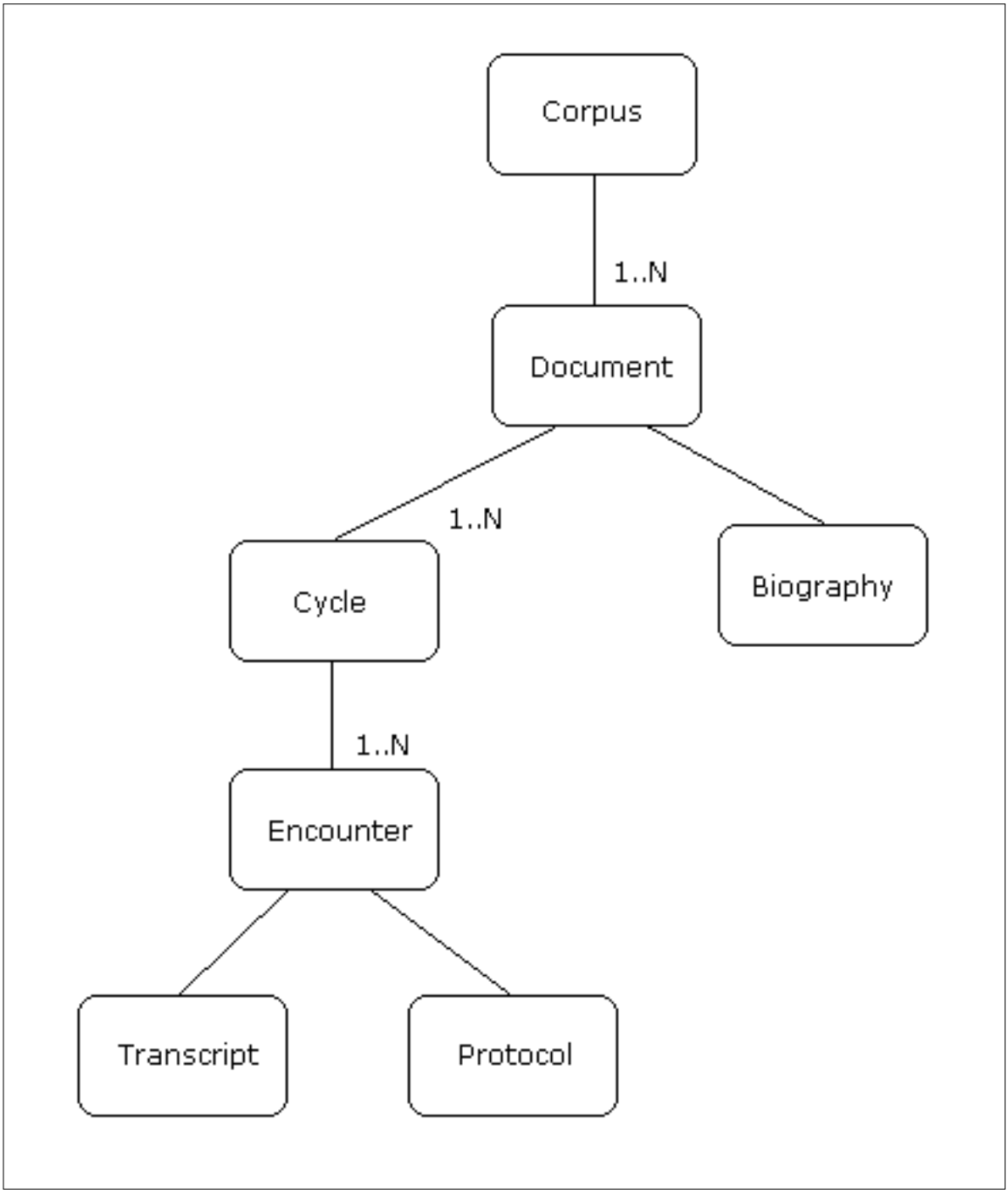
Introduction

The ESFSLD is a computerized archive of data collected by research groups of the ESF-project in five European countries: France, Germany, Great Britain, The Netherlands and Sweden. The project concentrates on the spontaneous second language acquisition of forty adult immigrant workers living in Western Europe, and their communication with native speakers in the respective host countries.

References

Feldweg, Helmut. 1992. The European Science Foundation Second Language Databank

Corpus Structure



Corpus Information

The corpus consists of a directory containing the names of the target languages. The target language directories have subdirectories with names of source languages in which the data files from the subjects are stored. A custom coding convention is used to classify and identify these files.

Document Information

Document data is distributed over the following three file types:

- RAW data files : the raw transcripts of the encounters
- PRT data files : contains a sort of protocol of an encounter
- BIO data files : contains socio-biographical information about the informants (subjects)

Header Information

The ESFSLD headers in the RAW and BIO files are flat structures with attribute-value pairs. No descriptions for attribute-value pairs from the BIO files were available. In addition to the main attribute-value pairs the BIO files there are three sections: fixed characteristics, variable characteristics and a list of encounters. Those sections are not described here.

Metadata Overview

Encoded in the filename		
	Informant type	There are three types of informants: longitudinal, control and long residents
	Source language	**** see transcription ****
	Target language	**** see transcription ****
	Informant	**** see transcription ****
	Cycle number	Cycle to which a particular session belongs
	Sequence number	Sequence number of an encounter within a cycle
	Activity type	Activity type of the encounter
Transcription (RAW)	Contains information about the raw transcripts of the encounters	
	Filename	(external) name of the file
	Informant	One-letter abbreviation and pseudonym used for informant(s) in the file
	Interviewers	One-letter abbreviation and name (pseudonym) used for interviewer(s) in the file
	Subject	**** same as informant ****
	Source language	Source language of informant (native language)
	Target language	Target language of informant (language to be learned)
	Date	Date of encounter
	Cassette	Label of audio/video cassette used for recording of encounter
	Recording	**** same as cassette ****
	Episode	Short description of transcribed episode
	Comments	Any comments concerning the episode
	Keywords	Keywords concerning relevance of transcribed data for specific analysis
	Transcribed by	Name of transcriber of the data
	Revised by	Name of revisor of the transcription
	Checked by	**** same as revised by ****
	History	Records of changes applied to the file
Protocol (PRT)	Contains a sort of protocol of an encounter	
Socio-biographical (BIO)	Contains socio-biographical information about the informants (subjects)	
	Group	?

Subject	?
Source language	?
Target language	?
Date of Birth	?
Sex	?
Religion	?
Fixed Characteristics	?
Variable Characteristics	?
Encounters	?

Gesture Database (GDB)

Introduction	References	Corpus Structure	Corpus Information
Document Information	Header Information	Metadata Overview	

Last update: 27-Feb-2001

Introduction

The Gesture Database from the Max Planck Institute in Nijmegen consists of the video recordings of speech and gestures that spontaneously accompany speech, and the annotations regarding gesture and speech in the recording. The recordings were made in different cultures, including the Netherlands, Italy, the USA, Japan, Turkey, Australian Aboriginal communities, Mexico, Belize, and Ghana. Speech events are recorded that elicits spontaneous gestures, such as narration of traditional stories and autobiographical stories, description of the local environment, and route direction.

Document Information

Document data consists of the following two types:

- MediaTagger files: the annotation files created by MediaTagger, developed at the MPI in Nijmegen
- Media files: QuickTime "containers" for a MPEG and an audiofile, Cinepak files (containing both audio and visual information)

Metadata Overview

NOTE: The following metadata elements are not official. The exact definitions are under development.

Researcher Name	The name of the researcher who made the recording
Recording Date	The date of when the recording was taken
Project Name	The name of the project for which the recording was made
Video Identifier	A unique identifier corresponding to the actual video tape on which the recording was made
Start Position	Start position of video fragment
End Position	End position of video fragment
Participant's Abbreviation	A unique abbreviation of the participant's name (identifier?)
Participant's Name	The participant's name
Filesize	The size of the digital recording
Backup Tape Identifier	A unique identifier corresponding to the tape which was used to backup the original recording
Backup Date	The date of when the backup was made
Backup Comments	Any comments concerning the backup process

International Corpus of English (ICE)

Introduction | References | Corpus Structure | Corpus Information
Document Information | Header Information | Metadata Overview

Last update: 13-Sep-2000

Introduction

The International Corpus of English began in 1990 with the primary aim of providing material for comparative studies of varieties of English throughout the world. Twenty centers around the world are preparing corpora of their own national or regional variety of English.

References

International Corpus of English Corpus Utility Programme (ICECUP). Information about metadata comes from the ICECUP online help manual (section: sociolinguistic variables).

Corpus Structure

The corpus is divided in several categories
(see http://www.ucl.ac.uk/english-usage/ice/design.htm)

Metadata Overview

Text code	The code number indicates the text category of the text within the hierarchy of the corpus as a whole
Speakers per text	The number of speakers, including extra-corpus speakers, in dialogues
Speakers Age	Age or age range of a speaker or author. The following ranges are used: 18-25, 26-45, 46-65, 66+
Speakers Gender	Speaker or author's gender
Speakers Education	Contains the speakers education (secondary, university). Secondary applies to those who have completed at most second-level schooling. University applies to those who have completed a course of tertiary education, though not necessarily at a University per se
Speakers role	The communicative role of a speaker in an exchange, e.g. interviewer, chairman
Medium	Indicates whether a broadcast was transmitted via TV or radio
Scope	The scope of a newspaper, e.g. local, national
Frequency	The frequency of publication of a newspaper, e.g. daily, weekly
Circulation	An approximate circulation figure for a newspaper at the time of the publication of the text

Linguistic Data Consortium Catalog (LDC)

Introduction	References	Corpus Structure	Corpus Information
Document Information	Header Information	Metadata Overview	

Last update: 13-Nov-2000

Introduction

The Linguistic Data Consortium supports language-related education, research and technology development by creating and sharing linguistic resources: data, tools and standards. The LDC's Catalog contains 168 corpora of language data.

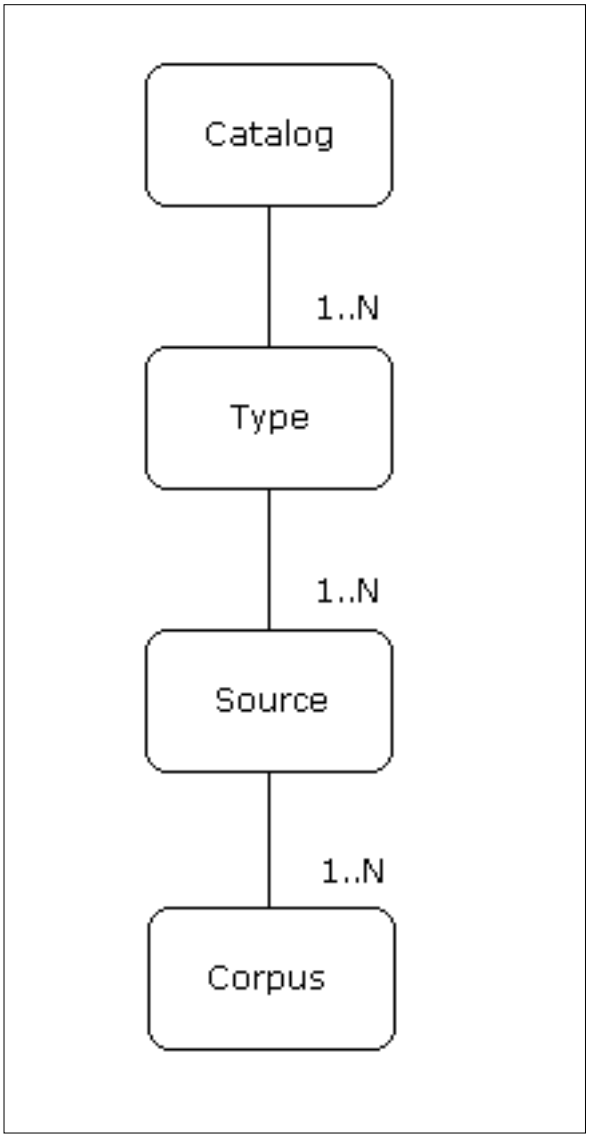
References

The Linguistic Data Consortium Catalog

LDC List of Catalog Fields (not used for this overview)

Catalog Structure

The catalog is an access structure on top of corpora where the metadata is about the corpora in the catalog. Corpora are first divided into major categories according to the type of data they contain, and then are further broken down into minor categories based on the source of the data.
(See http://morph.ldc.upenn.edu/Catalog/by_type.html)



Meta Date Overview

Catalog number	Contains a unique LDC catalog number
Name	Contains the name of the corpus
ISBN	Contains the ISBN
Data Sources	Contains the corpus data source (broadcast, conversation, microphone etc.)
Research Project	Contains the projects in which the corpus was used
Recommended Application	Contains the recommended applications for which the corpus is useful
Language	Contains the language used in the corpus
Membership Year	Contains the year in which the corpus was released
Corpus Type	Defines the type of the corpus (Lexicon, Speech or Text)

Multimedia Content Description Interface (MPEG-7)

Introduction	References	Corpus Structure	Corpus Information
Document Information	Header Information	Metadata Overview	

Last update: 09-Oct-2000

Introduction

The Multimedia Content Description Interface (MPEG-7) is an ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) standard developed by MPEG (Moving Picture Experts Group). MPEG-7 aims to create a standard for describing the multimedia content data that will support some degree of interpretation of the information's meaning, which can be passed onto, or accessed by, a device or a computer code. MPEG-7 is not aimed at any one application in particular; rather, the elements that MPEG-7 standardizes shall support as broad a range of applications as possible.

Only some elements from the Multimedia Description Schemes (MDS) part of MPEG-7 which we think are relevant for language resources are described in the overview.

References

The following sources are used to obtain the metadata information:

- Overview of the MPEG-7 Standard
- Multimedia Description Schemes XM from the MPEG working documents

Metadata Overview

Description of the media				
	MediaFormat DS			
		FileSize	The size, in bytes, of the file where the media profile is stored	
		Medium	The physical storage medium on which the media profile is stored (e.g., tape, CD, DVD)	
	MediaCoding DS			
		CompressionFormat	The compression standard used in the coding of the AV content (e.g., MPEG-1, MP3, JPEG).	
	MediaInstance			
		InstanceLocator	The location of the media instance. It is either a URL or a string for AV content not available on-line	
Description of the content creation & production				
	Creation DS	Describes the creation of the content, including places, dates, actions, materials, staff (technical and artistic) and organizations involved		
		Title DS	The title(s) of the AV content. Multimodal titles (of multiple types) may be defined: text, image, video, and audio	
			TitleText	The (textual) title of the AV content. The language in which the title is given is specified using an attribute
		Abstract	This annotation is aimed to provide a normalized way to access to a simple description of the AV content	

		Creator	The creator(s) of the AV content. It allows to describe individuals, organizations, groups, etc. Involved in the creation	
			CreatorType	Describes the creator of the AV content. It allows to describe individuals, organizations, groups, etc., involved in the creation
			Role	The role played in the creation (e.g., director, presenter, actor, contributor, author)
		CreationCoordinates	The place and date where the content was created. It comprises a locations and/or a date	
			CreationLocation	The place where the content was created
			CreationDate	The date when the content was created
		CreationMaterial	The devices and settings used for the creation of the content	
			DeviceInstrument	Device or instrument used for the creation (e.g., lens, films, piano)
			DeviceSettings	Setting of the device or instrument
	Classification			
		ClassificationType	DS describing the classification of the AV content	
		Country	The country code using ISO 3166-1 from where the AV content comes. It may be different than the location where it was created	
		Language	The language of the AV content	
		Genre	The genre that applies to the AV content. It can specify styles for a specific genre (e.g, Contemporary Classical Music)	
		Subject	The subject that applies to AV content. The subject classifies AV content from a point of view of types of program, without considering genre classification	
Description of the content usage				
	Rights	Who owns the AV content, and how the AV can be used		
		RightsType	Description of a link to the right holders and to the access rights information. It contains a unique reference to an identifier under management by an external authority	
	UsageRecord	Description of the past use of the AV content		
		Distributor	The distributor of the AV content in the concrete use described in the UsageRecord DS instance. Usually it will be an organization	
		Financial - Cost	Description of the cost associated to the creation and usage the AV content. Being derived from the InternationalPrice complex type, it expresses the cost with currency and value	
			InternationalPrice	
				InternationalPriceType: Description of a price
				Currency: The currency (using ISO 4217) of the price
				Value: The value of the price

Spoken Dutch Corpus (Corpus Gesproken Nederlands - CGN)

Introduction	References	Corpus Structure	Corpus Information
Document Information	Header Information	Metadata Overview	

Last update: 27-Feb-2001

Introduction

The Spoken Dutch Corpus Project is aimed at the construction of a database of contemporary standard Dutch as spoken by adults in the Netherlands and Flanders. Upon completion, the corpus will contain approximately ten million words, two thirds of which originate from the Netherlands and one third from Flanders. The Spoken Dutch Corpus comprises a large number of samples of (recorded) spoken text. In all about 1,000 hours of speech.

References

The Spoken Dutch Corpus (CGN-project)

Metadata Overview

Corpus Header	contains general information about the project and/or information which is equal for all samples				
	***type** *	describes to which type of document the header is part of (CORPUS)			
	***creator** *	name of the (final) producer of the header.			
	***version** *	version of the header that is adapted.			
	***update** *	gives the date of when the header is last modified			
	fileDesc	?			
		titleStmt	Information about the contents of the corpus		
			title	?	
			respStmt	?	
				respType	describes the task for which somebody/institute was responsible
				respName	name of the responsible institute
		editionStmt	number of the release		
			***release** *	?	
			***version** *	?	
		extent	size of the corpus		
			wordCount	total amount of words in the corpus	
			secCount	total amount of seconds of the corpus	
			byteCount	total amount of bytes comprising the corpus	

		extNote	additional information about the kind of counting(s). e.g. about the punctuation marks.	
	tempoAv	gives the average speed of speaking in the corpus		
		***wph** *	average amount of words per hour	
		***(id)** *	id of the discerning component	
	publicationStmt	information about the publication and distribution of the corpus		
		distributor	name of the distributor	
		pubAddress	address of the distributor	
		telephone	telephone number of the distributor	
		fax	fax number of the distributor	
		eAddress	email address of the distributor	
		availability	distributionregion of the (actual version of) the corpus	
			***region** *	
			***status** *	
		pubDate	date of distribution	
		copyright	name of the copyright holder	
encodingDesc	Documents the relationship between the texts and the sources			
	projectDesc	Description of the CGN project		
	samplingDecl	Description of the sampling method		
	editorialDecl	Information about the state of affairs during digitisation and annotation of the text		
		transduction	Describes the digitization and transcription process of recordings	
		segmentation	Describes segmentation principles in the corpus, e.g. division by speakers, utterances, sentences, words etc.	
	refDecl	Explains how (parts of) fragments are named and how they relate		
	classDecl	Description of the classification of the samples in the corpus		
	1..N	category	?	
			catDesc	?
profileDesc	Specific information about the corpus			
	langUsage	describes the language (variety) which is included in the corpus		
	wsdUsage	Contains one or more <writingSystem> sub-elements		
	1..N	writingSystem	indicates which ISO characterset is used.	
revDesc	Documents the applied changes			
1..N	change	?		
		date	date on which the update was made	
		respStmt	?	
			respType	description of the task
			resp	description of the change
			respName	name of the responsible institute

Text Header	?					
	***type** *	describes to which type of document the header is part of (TEXT).				
	***creator** *	name of the (final) producer of the header.				
	***version** *	version of the header that is adapted.				
	***update** *	gives the date of when the header is last modified				
	fileDesc	contains a (bibliographic) description of the corpus				
		titleStmt	information about the contents of the fragment			
			title	?		
			respStmt	?		
				respType	Describes the task for which somebody/institute was responsible	
				respName	Name of the responsible institute	
		extent	size of the fragment			
			wordCount	total amount of words in the fragment		
			secCount	total amount of seconds of the fragment		
			byteCount	total amount of bytes comprising the fragment		
			extNote	additional information about the kind of counting(s). e.g. about the punctuation marks.		
		tempoAv	average speed of speaking			
			***wph** *	average amount of words per hour		
		publicationStmt	information about the distribution of the fragment			
			distributor	name of the distributor		
			availability	spreading of the fragment		
				corpus	?	
				cd	?	
				date	?	
		sourceDesc	bibliographic description of the source			
			biblStr	bibliographic description		
				author	first initial(s) and last name (of the writer)	
				title	title	
				pubName	name (of the publisher)	
				pubPlace	place (of publishing)	
				pubDate	year of distribution of the used print	
			rec	?		
				***date** *	date of recording	
				***time** *	time of recording	
			source	indicates from where the material originates		

		producent	producer of the recording	
encodingDesc	Documents the relation between texts and sources			
	editorialDecl	Provides details of editorial principles and practices applied during the encoding of a text		
		correction		
			***type** *	?
			***status** *	checked YES/NO
profileDesc	Specific information about the corpus			
	textClass	indicates which classifications are relevant for the text		
		catRef	?	
			***target** *	one or more catDesc values for the fragment
			keywords	keyword chosen from a limited list
			***term** *	?
	particDesc	?		
		person	?	
			***id** *	speaker identification code
			***role** *	speakers role
			***age** *	interpretation of the age of the speaker during the recording
		interaction	interaction between participants	
			***type** *	?
			***active** *	amount of active (identified) speakers
			***passive** *	amount of passive (unidentified) speakers
		relation	relation between the speakers	
			***active** *	speaker identification of the active speaker in a directional relation or all speakers in a non-directional relation
			***desc** *	description of the relation
			***mutual** *	indicates whether the relation holds for all speakers or is directional
	settDesc	?		
		region	province where the recording is taken	
		locName	place where the recording is taken	
		locale	description of the space where the recording is taken	
		activity	describes in short what the speakers are doing	
	recCondition	?		
		recMedium	?	
			***type** *	medium of recording

				microphone	type of microphone used to make the recording	
				micDistance		
					person	speaker ID
					dist	?
					cm	distance in centimeters
				noise	description of the background noise with the recording	
			digitisation	?		
				opname	analog / digital ?	
				verwerking	analog / digital ?	
				status	analog / digital ?	
		revDesc	Documents the changes that are applied			
		1..N	change	?		
				date	Date on which the update was made	
				respStmt	?	
				respType	description of the task	
				resp	description of the change	
				respName	name of the responsible institute	
Participant Header	?					
	***type** *	describes to which type of document the header is part of (PARTICIPANT).				
	***creator** *	name of the (final) producer of the header.				
	***version** *	version of the header that is adapted.				
	***update** *	gives the date of when the header is last modified				
	particDesc	Gives general information about the speaker				
		person	?			
			***id** *	speaker identification code		
			***sex** *	speaker's gender		
		birth	?			
			***year** *	speaker's year of birth		
			***place** *	speaker's place of birth		
			***reg** *	region where the speaker is born		
		language	?			
			firstLang	language variant in which the speaker is raised		
				***lang** *	?	
				***dialect** *	?	
			homeLang	language variant the speaker uses at home		
				***lang** *	?	

		***dialect** *	?
	workLang	language variant the speaker uses at work	
		***lang** *	?
		***dialect** *	?
residence	?		
	***place** *	speaker's place of residence	
	***reg** *	the region where the speaker is living	
	***size** *	indication of the size of the population where the speaker's living	
education	?		
	***place** *	place where the speaker followed his/her education	
	***reg** *	region where the speaker followed his/her education	
	***opleiding** *	highest education the speaker finished	
	***level** *	level of education	
occupation	?		
	***job** *	speaker's job	
	***level** *	job level indication	
notes	Other remarks concerning the speaker, e.g. participation in other projects, other places of residence, etc.		

Language Engineering Resources

Last updated: 27-Feb-2001

The following resources are searched for metadata. Resources marked in GREEN are included in the overviews.

Standards

- CES
- Dublin Core
- EAD
- Harmony
- MATE
- MPEG-7
- TEI
- Tipster
- Open Archives

Corpora

- ASEDA
- ATLAS
- BLC Collections
- BNC
- CDEL
- CHILDES
- ELRA Catalog
- ESFSLD
- Gesture DB
- ICE
- LACITO
- LDC Catalog
- MICASE
- NEGRA
- SCOIL Catalog
- SDC (CGN)
- UHLCS

Tools / Systems

- Alembic Workbench
- Browsable Corpus
- CSLU
- GATE
- GSearch
- HIAT
- SignStream
- SIL Tools
- Transtool

References

- ANLC
- Linguistic Annotation

(Overview from LDC)

Overview Format

Introduction

Gives a short description of the corpus or metadata standard.

References

Lists all the references which are used as an information source for the contents of the overview

Corpus Structure

Describes the overall structure of the corpus. A diagram can be included to visualize the structure.

Corpus Information

Gives information about the corpus as a whole.

Document Information

Gives information about the documents contained in the corpus.

Header Information

Gives information about headers from which the metadata elements are extracted.

Metadata Overview

Contains the overview table of metadata elements together with their definitions. The structure is maintained by increasing the indent for each lower level. Colors are used to code the following:

White	Identified metadata element. This element is also included in the global metadata overview
Green	Identified grouping. Labels a group of elements from a lower level
Gray	Possible metadata element. This element is not yet included in the global metadata overview.
Brown	Used to specify the amount of elements. E.g. "1..N" indicates that there can be 1 to N elements.