

A Browsable Corpus: accessing linguistic resources the easy way.

D.G. Broeder, H. Brugman, A. Russel, P. Wittenburg.

Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen. The Netherlands.
Email: broeder@mpi.nl

Abstract

Recently the Browsable Corpus concept was introduced at the Max Planck Institute for Psycholinguistics. Generalization of this concept could help resource discovery and access for Linguistic Resources.

1. Introduction

A large number of linguistic resources are being generated across the world for a variety purposes in science and language engineering and it is impossible to find – let alone access - most of them. A few consortia like ELRA and LDC are collecting and cataloguing some linguistic resources, but most resources fall outside the ambit of these catalogues, even though the data the resources contain might be useful in ways not anticipated by the original collectors.

To make resource discovery easier, we propose to define an universe of meta-descriptions for linguistic resources. The meta-descriptions will incorporate links to one another, thus forming a structure that can be browsed and searched.

A scheme like this has been introduced at the Max Planck Institute for Psycholinguistics (MPI) and has been applied to a variety of language resources (LRs) from a variety of researchers and research areas. Although not all MPI's LRs have been incorporated in this way, enough has been done to suggest that this approach shows promise [1].

A joint NSF/EC initiative [2] is also working on meta-data description for LR's.

2. Structuring with meta-data

At the Max Planck Institute for Psycholinguistics (MPI) the concept of the Browsable Corpus (BC) was introduced to help organise and structure a growing mass of multi-media language resources (LR). The BC scheme depends on the creation of an accompanying meta-description file (MD-file) for each individual LR and every LR bundle, e.g. a transcription, or transcription plus media files for a multi media corpus. Those MD-files

called session MDFs contain a variety of information on the LR's and also specify where to find the individual resources. Above this lowest layer (of meta-description files that point to the LRs) a hierarchy of layers of other meta-descriptions are built up. Each MD-file refers to at least one MDF from a lower layer, so that the MD-files form subsets that share certain meta-data attributes with the session MDF. Together the MD-files form a pyramid culminating in a top MDF that forms the entry point of the corpus universe (see Figure 1).

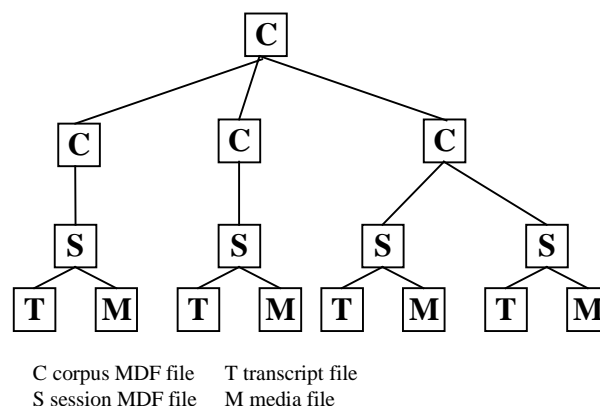


Figure 1

The structure created by the linked MDFs creates a space that can be navigated by a browser tool specifically designed for this task. At each node within this space the browser will display the meta-data for the selected MD, in a way that is closely analogous to a web browser surfing the World-Wide-Web (see Figure 2). Since the links between the MDs and LRs are specified as URLs and the browser can access MDFs via the HTTP protocol, the MDFs and LRs may be distributed over different physical sites, reinforcing the WWW analogy.

even a home PC, and does not want to depend on special

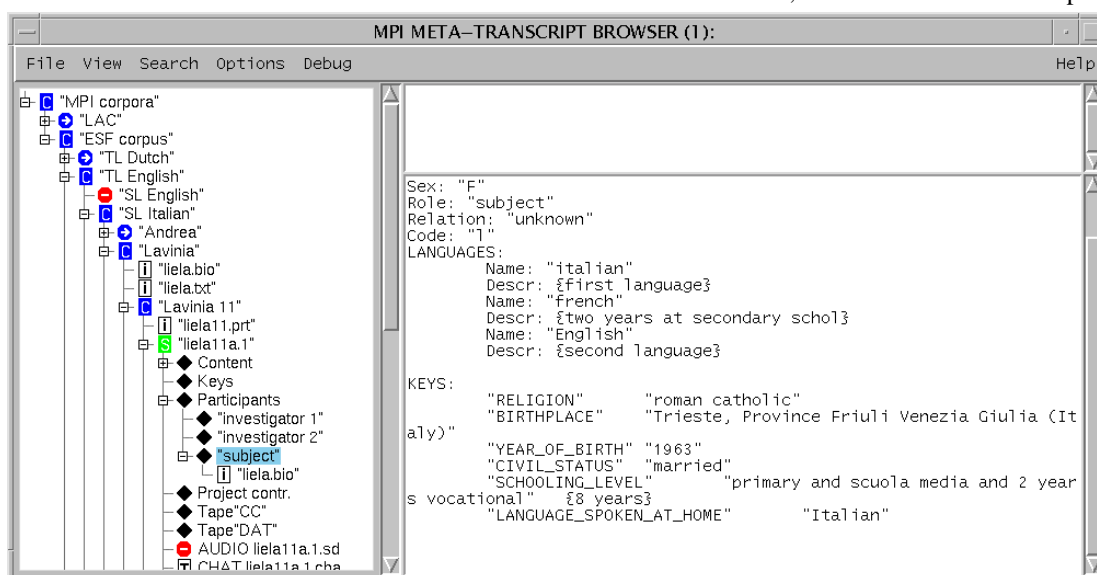


Figure 2

3. Tools

As stated above, a special browser tool was developed to navigate the universe of linked meta-descriptions. Originally it was based on the Tcl/Tk scripting language for platform independence, although now we are finalizing a JAVA version. The browser parses the XML meta-description format (see The structure of the Meta-Description File) and displays the corpus structure and meta-data elements.

At the moment the BC browser is a separate application and although it can work with a standard web browser such as Netscape to display HTML data it is not part of Netscape itself. The possibility of making the BC browser a JAVA applet that can be automatically downloaded via Netscape, and run within it, is now under investigation. This would make the task of supplying new versions considerable easier.

The browser tool is coupled with specific tools that work with individual LRs so that the user may start up such a tool by simply clicking on a transcription file node or a media file node in the browser. Of course, for this to work, there has to be an appropriate coding scheme for LRs. For the MPI's BC browser we use an extension of the mime-type scheme. For accessing the LR's remotely over the internet, in the same way that the BC browser tool accesses the MD-files, the tools must be able to use an appropriate protocol such as HTTP.

In the simplest form of BC, the meta-data associated with the LRs is stored in simple files, and not in the database structures used by some corpora. This is done for the benefit of the researcher who wants to store (part of) a BC-description and resource on a local workstation or

remote servers. Researchers should not feel that coupling their data into the BC universe requires them to relinquish control of their data.

This does not mean that meta-data stored in files is the only possibility. Meta-description files could be generated in real time by a program working on data extracted from a data-base, or the meta-description files themselves could be stored in a data-base. The browser does not care about the origin of the XML data presented by the remote HTTP server.

The whole BC scheme opens up vistas of interesting possibilities. It has been suggested, for instance, that it could support an infrastructure where researchers who reference part of a browsable corpus in their publications could leave a visible citation mark attached to a BC node.

4. The structure of the Meta-Description File

As explained above, the MDF files contain meta-data information about the LR at the leaves of the MDF tree, plus pointers to the LR's and other MDF files. XML [3] was chosen as the format for the MDF because of its power to express structure and the opportunity it gives us to use the expanding range of XML capable software now available.

The meta-data information is partitioned in a number of structures being: "General", "Content", "Participants" and "Files". The "General" structure contains general information on the LR pointed to such as "Project controller", "Access rights" information etc. The

“Content” structure provides information like the language spoken, the nature of the linguistic interaction e.g. dialogue/monologue, spontaneous or prepared text. The “Participants” structure gives biographical information on the participants, their linguistic status and describes their mutual relations. The information on the coding and location of the LR’s themselves is contained in the “Files” structure. A complete description of the MDF is given by the DTD and an example of part of an existing MDF file is given in Appendix A.

At the moment the DTD we use is specifically tailored for use with corpora at the MPI and it is unlikely that it will satisfy the needs of a broader community. Therefore we expect to have to adapt the DTD or create new DTDs that will allow the modeling of meta-data appropriate to other language resources such as lexica.

A limitation of DTD’s, when they are used to specify XML documents, is the lack of instruments to impose constraints on the data -- for example to say that the value of the “AGE” attribute must always have a positive value. There have been proposals submitted to W3C for an alternative way of specifying the structure of an XML document based on XML schema’s [4]. We will adopt XML schemas as soon as reliable parsers implementing the complete XML schema definition become available.

So far the biggest problem with the introduction of BC concept at the MPI has been the generation of the MD files. In cases where existing corpora include meta-data in a well defined form, the problem can be solved by automatically transforming this “legacy meta-data” into the new MDF format. When this is not the case we have used “knowledgeable persons” filling in templates to create new MD files. It soon became clear this was not the most productive approach and we have now developed an editor with which the XML-naïve researchers can generate the MD files themselves.

5. Searching resources

Using MDF’s to structure corpora not only facilitates browsing the corpus and associates tools with specific resources, it also facilitates searching. For example if we were interested in knowing about all the transcriptions in a corpus where the speaker is above 40 years of age, we could use a search tool that ran down the MDF hierarchy and selected all the transcriptions that matched that criterion. We can of course only search for items that are available in the set of meta-data items and have been coded in a consistent format.

In its simplest form a search engine would act just like a web-crawler going through the hierarchy of meta-description files checking for meta-data items that match the search criterion. This would not be an acceptable approach for a site with many meta-description files, where this would simply take too long, even if all the

meta-description files were “virtual” files checked out of a data-base. In that case a meta-data search would do better to query a data-base directly. The simple “crawler” approach should always be available, to cater for those sites that did not have the resources to set up and maintain a data-base.

An interesting concept is the possibility of using the search results to dynamically create a new hierarchy of meta-descriptions. A user could select a node in a BC hierarchy and then specify that the next division should be for instance, be between speakers above and below a certain age, dividing the meta-description nodes under the selected nodes into the two subsets specified.

Although these are all interesting ideas for the future, at the moment all we can do are some simple searches with Perl scripts from outside the BC browser application. We hope to realize an integrated solution before the end of this year.

6. Towards a Browsable Corpus Universe of Linguistic Resources

The successful introduction of our local version of Browsable Corpus suggested that we should put some effort into creating a distributed universe of linguistic resources. We have developed an MDF browser that can access MDF’s via the internet and have developed and are developing tools that are also able to access the LRs via the Internet. To demonstrate this we are finalizing the connection of the browser to the EUDICO system [5] that will be the basis for the corpus exploitation software for a.o. The Spoken Dutch Corpus [6].

If we can establish a suitable portal we could access LRs across the world using existing software. As mentioned above, the DTD we now use for the MDFs is specifically tailored for use with corpora at the MPI and we know that it should be generalised, or new DTDs (or XML schemas) should be added to allow other LRs to be modelled. We hope that the EC/NSF initiative [2] will inform the generalisation of this DTD and the generation of new DTDs or XML schemas.

7. References

- [1] MPI’s BC website reference
<http://www.mpi.nl/world/tg/lapp/browscorp/browscorp.html>
- [2] International Standards in Language Engineering – a EC 5th Framework and NSF initiative
- [3] <http://www.w3c.org/XML/>
- [4] <http://www.w3c.org/XML/Schema.html>

[5] A. Russel, H. Brugman et al. The EUDICO project, multi media annotation over the Internet. To be published in the LREC2000 proceedings. See also <http://www.mpi.nl/world/tg/lapp/eudico.html>

[6] Oostdijk. N. 2000. The Spoken Dutch Corpus, Overview and first evaluation. To be published in the LREC2000 proceedings.

8. Appendix A

An example of an XML Meta Description file

```
<?xml version='1.0'?>

<!DOCTYPE METATRANSCRIPT SYSTEM
"metatranscript.dtd" [

<!ENTITY % esf-config SYSTEM "esf-config.xml">

%esf-config;
<!ENTITY datadir "&english-
dir;/longitudinal/italian/liela">
]>

<METATRANSCRIPT NAME="liela17k.1.xml"
DATE="19990501" VERSION-ID="0" GEN="rs"
PARENT="%xmldir;/liela17.xml">

<SESSION NAME="liela17k.1" LEVEL="0" DATE="29-
FEB-1984">

<DESCRIPTION> This is the ESF subcorpus TL
English, SL Italian, subject
Lavinia, cycle 1, sequence 7</DESCRIPTION>

<FDESCRIPTION FORMAT="text/html"
SRC="%&htmldir;/liela17.html" />

<ACCESS DATE="19981023"> Free </ACCESS>

<PROJECT_CONTROLER NAME="ESF" />

<CONTENT_KEYWORDS=" " >

<LANGUAGE TAG="Language Spoken" NAME="English" >
<DESCRIPTION>Target Langage</DESCRIPTION>
</LANGUAGE>

<DESCRIPTION>(Ep 2) 033-083 At the bank. L asks
the exchange rate of the pound against the lira.
General conversation whilst walking to the bank.
(Ep. 6) 231-At the travel agents. (Because of a
misunderstanding-as it was an agency dealing with
flights-the clerk took quite a long time and the
tape ran out. Therefore the end of the task is
not recorded.) SIDE TWO (Ep. 7) 002-036 L talks
to margaret about the travel agency. Discussion
on where to ask for the way to Boots.
</DESCRIPTION>

<INFOFILE NAME="liela17.prt"
SRC="%&datadir;/liela17.prt" FORMAT="text/text">
  <DESCRIPTION> The protocol file gives
information about the
whole encounter cycle 1 sequence 7</DESCRIPTION>

  </INFOFILE>
</CONTENT>

<PARTICIPANTS>
<PERSON NAME="investigator 1" FULLNAME="margaret
simonot" SEX="unknown" ROLE="investigator"
CODE="m " AGE="unknown" RELATION="unknown">
</PERSON>
```

```
<PERSON NAME="investigator 2" FULLNAME="elisa
sponza" SEX="unknown" ROLE="investigator" CODE="e
" AGE="unknown" RELATION="unknown"> </PERSON>

<PERSON NAME="subject" FULLNAME="Lavinia" SEX="F"
ROLE="subject" CODE="1" AGE="see Keys"
RELATION="unknown"> <LANGUAGES>
  <LANGUAGE NAME="italian" >
    <DESCRIPTION>first language</DESCRIPTION>
  </LANGUAGE>
  <LANGUAGE NAME="french" >
    <DESCRIPTION>two years at secondary
schol</DESCRIPTION>
  </LANGUAGE>
  <LANGUAGE NAME="English" >
    <DESCRIPTION>second
language</DESCRIPTION>
  </LANGUAGE>
  <DESCRIPTION>source language Italian,
target language English</DESCRIPTION>
</LANGUAGES>

  <KEYS>
    <KEY NAME="RELIGION" VALUE="roman catholic"
  > </KEY>
    <KEY NAME="BIRTHPLACE" VALUE="Trieste,
Province Friuli Venezia Giulia (Italy)" > </KEY>
    <KEY NAME="YEAR_OF_BIRTH" VALUE="1963" >
  </KEY>
    <KEY NAME="CIVIL_STATUS" VALUE="married" >
  </KEY>
    <KEY NAME="SCHOOLING_LEVEL" VALUE="primary
and scuola media and 2 years vocational" >
  <REMARK>8 years</REMARK></KEY>
    <KEY NAME="LANGUAGE_SPOKEN_AT_HOME"
VALUE="Italian" > </KEY>
  </KEYS>
  <INFOFILE NAME="liela.bio"
SRC="%&datadir;/liela.bio" FORMAT="text/text" />
</PERSON>

</PARTICIPANTS>

<TAPE ID="itlg8.7" POSITION="033-036" FORMAT="CC"
>
<DESCRIPTION> This is the original recording
</DESCRIPTION>
</TAPE>
<TAPE ID="liela17" POSITION="unknown"
FORMAT="DAT">
<DESCRIPTION> This is a copy from the CC
tape</DESCRIPTION>
</TAPE>

<KEYS>
<KEY NAME="SOUND_LINKING" VALUE="??" >
<REMARK></REMARK></KEY>
</KEYS>

<FILES>

  <TRANSCRIPTION-FILE
SRC="%&datadir;/liela17k.1.cha"
FORMAT="text/chat">
  <REMARK>Was generated from ESF
transcript</REMARK>
  </TRANSCRIPTION-FILE >
  <TRANSCRIPTION-FILE
SRC="%&datadir;/liela17k.1.tr" FORMAT="text/esf">
  <REMARK>Original ESF trancript</REMARK>
  </TRANSCRIPTION-FILE >
```

```
<MEDIA-FILE SRC="&datadir;/lielal7k.1.sd"
FORMAT="audio/esps" START="unknown"
      DURATION="unknown" AUDIO-
QUALITY="unknown">
  <REMARK></REMARK>
  </MEDIA-FILE>

</FILES>

</SESSION>

</METATRANSCRIPT>
```