Multilingual Corpora for Cooperation Final Report 6 October 1995.

(subcontract to EU project LRE 61-101 "International Co-operation for EAGLES")

Susan Armstrong[b], Masja Kempen[a], David McKelvie[a], Dominique Petitpierre[b], Reinhard Rapp[b], Henry Thompson[a]

[a] Human Communication Research Centre, Edinburgh University,
2 Buccleuch Place, Edinburgh EH8 9LW,
SCOTLAND
Tel:+44 131 650-4630
Fax:+44 131 650-4587
email: dmck@cogsci.ed.ac.uk

[b] Institut Dalle Molle pour les Études Sémantiques et Cognitives (ISSCO)
54 route des Acacias, Genève,
SWITZERLAND
Tel: +41 22 705 7113
email: susan@divsun.unige.ch

Contents

Chapter 1

Introduction

1.1 Background

We report here on the corpus data acquired and prepared under the "MLCC " project (Multilingual Corpora for Cooperation). This project was elaborated within the framework of the LRE (Linguistic Research and Engineering) under the program for "International Cooperation for Eagles" (LRE 61-101). The goal of the MLCC project was to produce a multilingual corpus with two main components, a polylingual document collection of comparable material and a parallel corpus of translations. The data is destined to be published on CD-ROM and distributed at cost with minimal restrictions.

The MLCC corpus is intended to answer the needs of the European research community for comparable data in a wide range of languages. These resources will serve as a basis for new technology development and ultimately contribute to the realization of better commercial linguistic products. It has become clear that large amounts of data are necessary to assure the development of wide-coverage, robust and effective applications. The polylingual document collection, consisting of similar documents in six languages, provides an important addition to monolingual collections and will assure that researchers can carry out similar studies in their own language. This collection, with obvious connections to the TREC/Tipster materials used in the United States, will enable comparability in research on an international scale. The material can also provide the basis for European evaluation programs in Information Retrieval and NLP. The parallel data consisting of translated data in nine languages provides a wealth of material for translation studies as well as new technology development.

In this report we document the acquisition and preparation of the data delivered in fulfillment of the MLCC contract. We begin with a general overview of the data, followed by a detailed description of the two main components of the corpus.

1.2 Description of the corpus

The MLCC text corpus has two main components – one set to allow comparable studies to be carried out in different languages and one set as the basis for translation studies. We refer to the first set as the Polylingual Document Collection, a collection of newspaper articles from financial newspapers in 6 languages (Dutch, English, French, German, Italian and Spanish). The Polylingual Document Collection consists of the following sub-corpora:

- Dutch - *Het Financieel Dagblad* - 1992-1993
- English - *The Financial Times* - 1993
- French - *Le Monde* - 1992-1993
- German - *Handelsblatt* - 1986-1988
- Italian - *Il Sole 24 Ore* - 1992-1993
- Spanish - *Expansion* - 1994

The second set is a Multilingual Parallel Corpus, consisting of translated data in nine European languages. The languages are Danish, Dutch, English, French, German, Greek, Italian, Portugese, and Spanish. The parallel data, provided by the European Commission, comprises two sub-corpora from the Official Journal of the European Communities.

- Official Journal of the European Commission,
  C Series: Written Questions 1993
- Official Journal of the European Commission,
  Annex: Debates of the European Parliament 1992-1994

The choice of data was motivated by the multilingual needs of the European community with special attention to translation concerns. On the practical side, these texts were identified on the basis of the feasability of acquiring and preparing such data within the constraints of the project. This corpus will by no means satisfy the need for very large amounts of varied resources in many languages, though it does represent a modest first step.

1.3 Polylingual Document Collection

This part of the project was directed at collecting comparable financial journalism material from six EU languages (Dutch, English, French, German, Italian and Spanish) from a common time period.

Initially, we created a list of possible newspapers for each language. Having chosen the newspapers, we then made contact with the newspapers asking if they would be prepared to let us have electronic copies of past issues. The newspapers that we contacted were generally helpful and generously agreed to our requests. We would like to thank the editorial boards of the newspapers noted below for their generosity in donating data to this project and their help in providing information about the data and its markup. However, negotiation of licences and delivery of data were very time-consuming activities. Once we had received the corpus data, SGML -markup proceeded smoothly as documented in following chapters.

1.3.1 Dutch - *Het Financieel Dagblad* - 1992-1993

The corpus contains articles from the Dutch financial newspaper *Het Financieel Dagblad* editions of 2^{nd} January 1992 through to 24^{th} December 1993. It contains around 8,5 million words of text. See chapter 2 for more details.

1.3.2 English - *The Financial Times* - 1993

The corpus contains articles from the British financial newspaper *The Financial Times* editions from the year 1993. The corpus contains around 30 million words. See chapter 3 for more details.

1.3.3 French - *Le Monde* - 1992-1993

A corpus of articles from the French newspaper *Le Monde*, consisting of two years worth (1992-1993) of articles on financial subjects, approximately ten million words. Edinburgh have a licence with the newspaper allowing us to distribute this corpus for research purposes. See chapter 4 for more details.

1.3.4 German - *Handelsblatt* - 1986-1988

This subcorpus consists of articles from the German financial newspaper *Handelsblatt* from the period 02.01.1986 to 15.06.1988. It contains some 33 million words. Unfortunately, the time period of these articles was not from the same (1992-1993) period as the others. It may be possible to obtain more recent articles from *Handelsblatt*. See chapter 5 for more details.

1.3.5 Italian - *Il Sole 24 Ore* - 1992-1993

The corpus described here contains articles from the Italian financial newspaper *Il Sole 24 Ore* from the year 1992. This corpus contains some 1.88 million words. The data was obtained from the newspaper by PISA University, who have a licence allowing re-distribution of the corpus. The SGML -markup was done by Edinburgh. See chapter 6 for more details.

1.3.6 Spanish - *Expansion* - 1994

This subcorpus contains articles from the Spanish financial newspaper *Expansion* editions of 21-10-91 through 24-10-91 and 14-05-94 through 27-12-94. It contains some 10 million words. See chapter 7 for more details.

1.4 Multilingual Parallel Corpus

This part of the project was directed at collecting parallel texts in the 9 European Union official (as of 1993) languages and consists of corpora from the Office of Publications of the European Community. Initial meetings were held with members of the Office of Publications to identify potential collections of parallel data for inclusion in the MLCC corpus. The two series agreed on were the "Written Questions" (from the EU Office of Publications) and the "Parliamentary Debates" (from the Printing Offices of Parliament). The former collection was quickly supplied by the Office of Publications from in-house resources and delivered to MLCC in Spring 1994. The collection of the Parliamentary Debates, however, proved somewhat more difficult (and time-consuming) given that the only copy of some of the data was held in four different European printing offices. In what follows, the two parallel collections are presented separately.

1.4.1 Written Questions

The first parallel corpus included in the MLCC collection consists of records of questions and answers regarding European Community matters. The data is published regularly as one section of the C series of the Official Journal of the European Community in all official languages (previously nine and currently, as of 1995, eleven languages). This corpus contains written questions asked by members of the European Parliament and corresponding answers from the European Commission in 9 parallel versions (languages: Danish, Dutch, English, French, German, Greek, Italian, Portuguese and Spanish). The total size of the corpus is approximately 10.2 million words (ca. 1.1 million words per language).

The data was acquired by ISSCO from the Office of Publications of the European Community (OPOCE) and consists of material published in 1993. The texts were prepared by LTG. See chapter 8 for a detailed report on this corpus.

A subset of the corpus of "Written Questions", acquired and prepared in 1994, was made available to the MULTEXT project as a test case for software and resource development.

We are still negotiating with (OPOCE) to allow the release of this data to a wider audience.

1.4.2 Parliamentary Debates

The second parallel corpus defined for the MLCC collection is the records of Parliamentary sittings published as an annex to the Official Journal of the European Community – *Debates of the European Parliament*.

The Parliamentary Debates are a record of what was said by members of the meeting as well as written input provided to the meeting. The original data from which the translations are produced consists of a transcript of the sittings, each member speaking in the language of his choice. This version is circulated to all speakers for possible revision before it is sent out for translation. The final version, as collected and prepared for the MLCC corpus, consists of nine parallel versions of the material. The texts delivered comprise the Debates of Parliament from January 1992 to July 1994. This sub-corpus contains some 5 to 8 million words per language. See chapter 9 for more details.

1.5 Status of the Deliverables

1.5.1 Licence agreements

Our intention is that the MLCC corpora will be made publicly available under the auspices of the new European Linguistic Resource Association (ELRA). Since the ELRA has only just been set up, we originally made temporary licence agreements with the data providers. The licences made with the *The Financial Times* newspaper are included as an example in appendix B. These agreements will need to be re-negotiated – to this end, contact with ELRA has been established.

1.5.2 Final Markup

A full list of the files in the MLCC corpus is given in Appendix C. The data has been checked for SGML conformance in according to the DTDs elaborated for the sub-corpora. The majority of the data is also TEI conformant as documented in Appendix A. The Parliamentary Debates sub-corpus is still in need of some work to achieve full TEI conformance. Prior to public distribution, a certain amount of automatic checking and hand correction would also be desirable. We estimate an additional 3 to 4 months work will be required to prepare the full corpus for CD-ROM publication.

Chapter 2

The Dutch *Het Financieel Dagblad* Corpus

2.1 Description of the corpus

Size The size of the *Het Financieel Dagblad* data is 60.6 megabytes and contains around 8.5 million words. The raw data are in the files `orig/Ed9*.Lst.gz`, `orig/DUTCH.HFD.gz`, and `orig/ENGLISH.HFD.gz`. The SGML marked-up versions are in the files `data/Ed9*.Sgm.gz`, `data/DUTCH.HFD.Sgm.gz`, and `data/ENGLISH.HFD.Sgm.gz`.

Time period The corpus contains articles from the *Het Financieel Dagblad* editions of 2^{nd} January 1992 through to 24^{th} December 1993.

2.2 Structure of the corpus

This corpus has been SGML coded using the *newspaper.dtd*. Each file of this corpus consists of a <DIV0> containing a number of <DIV1>s representing different articles from *Het Financieel Dagblad*.

Articles are of the following general structure:

```
            ART
             |
     headline + date
            / \
           /   \
        INDEX   TEXT
```

Articles start off with the headline and publication date. This is followed by an (optional) INDEX followed by the actual article itself – including the headline again.

The INDEX consists of a number of LISTs, which appear at the start of the vast majority of articles. They are part of bibliographic information. Each list contains a list of ITEMs which depend on the list type. See below for more details.

An article contains the following SGML structure:

HEAD - Headline of the article.

OPENER - Contains a DATELINE element which gives the date of publication of the article (in the form YY-MM-DD).

INDEX - (optional) Contains several LIST elements. These lists have a TYPE attribute with value *descriptor, trade, country, company* or *personname*. See below for more details.

TEXT - Contains the text of the article, which is structured as follows:

- **HEAD** - One or more heads. Heads have a TYPE attribute with value *headline* or *subheading*. The first occurrence of HEAD will be the headline and the rest will be subheadings.
- **BYLINE** - (optional) This contains the name of the agency or journalist responsible for the article and/or the place where it was written, coded as <NAME> elements that have a TYPE attribute with value *person* or *place* respectively. BYLINEs occur either after the initial HEAD or at the end of the article.
- **P** - A number of <P> elements containing the main text of the article divided into paragraphs.
- **Figures** - Some paragraphs may contain information about any photographs or graphs which appeared in the printed article, but have been omitted from this corpus.

ID attributes on the articles are the original identifiers given to the articles by the FD.

No sub-paragraph markup has been added, except that omitted figures have been coded as e.g

```
<FIGURE><HEAD>Op 9 oktober werd de eerste vestiging van Toys R
Us geopend in Arnhem. Filialen in Rotterdam, Muiden, Eindhoven
en Utrecht volgen.</HEAD>
<FIGDESC>omitted photo</FIGDESC> by: ANP</FIGURE>
```

2.2.1 Lists

These lists, which appear at the beginning of each article, are part of the bibliographic information attached to each article. They consist of a list of ITEMs which depend on the list type as follows:

descriptor - ITEMs are key phrases describing the contents of the article.

trade - ITEMs are names of industry/trade connected to the subject matter of the article.

country - ITEMs are names of geographic areas connected to the subject matter of the article.

company - ITEMs are the names of companies/organisations which are referred to prominently in the article.

personname - ITEMs are the names of people referred to in the article.

For the lists of type *descriptor*, *trade* and *country*, each item is provided with an identifying number. These numbers have been put as a reference inside the <ITEM>. An example follows here:

```
<LIST type=country>
<ITEM ref= 911.01 > Nederland </ITEM>
</LIST>
```

2.3 Structure of the original

The corpus is a concatenation of articles. The format of the Financieele Dagblad full text source data is shown below:

```
    Document 4 ^M
^M
TI  TOPMAN JOHN SCULLEY VERTREKT BIJ APPLE NA DRAMATISCHE WINSTVAL^M
PD  931019^M
DC  663 / 654.4 / 622 / 616 / 623 / 42 / 618 /^M
DE  tussentijdse mededelingen / winst / ondernemingsbestuur. management /^M
    reorganisatie / ondernemingsplanning / werkloosheid / concurrentie /^M
CC  913.01 /^M
CN  VS /^M
PC  836.99 /^M
PN  computers /^M
CO  apple /^M
NP  j. sculley / m. markkula / m. spindler /^M
TX  TOPMAN JOHN SCULLEY VERTREKT BIJ APPLE NA DRAMATISCHE WINSTVAL Nadat^M
    hij in juni al op een zijspoor was gezet, is president John Sculley^M
    ......
TX  Mike Markkula is inmiddels benoemd als de nieuwe bestuursvoorzitter.^M
    ......
TX  Sculley kwam in 1983 bij Apple terecht, nadat hij eerder de hoogste^M
    ......
```

Each article starts with a document number. This has been put inside <DIV1> as the value of attribute *n*. This is followed by:

TI - Headline

PD - Publication date

DC/DE - Descriptors and their corresponding codes

CC/CN - Countries and their corresponding codes

PC/PN - Trade and corresponding code

CO - Company

NP - Names of people

TX - Text
Each TX indicates the start of a new *paragraph*. The first occurrence of TX starts with capitalised text which is the headline of the article. The rest of capitalised text inside articles are usually *subheadings*.

2.3.1 Existing markup

The abovementioned markers are the only form of markup that the original data contained. These are sufficient to markup the basic structure of each document. With a few perl-programmes, these markers have been converted into SGML markup. The major restructuring needed to be done on the TX-sections; Capitalised text needed to be put inside HEADs, Names of authors and places inside BYLINEs, references to photographs inside FIGUREs.

All files have been checked for their conformance to the DTD. Because of inconsistency in the use of capitalised text in the original data, quite a lot of hand correction of headlines and sub-headlines markup was necessary.

Chapter 3

The English *The Financial Times* Corpus

3.1 Introduction

The *The Financial Times* have provided the MLCC project with a large corpus of articles from their archives, which is described below.

Size The size of the *The Financial Times* data and contains around 30 million words. The raw data are in the files `data/orig/tape*.gz`. The SGML -marked-up versions are in the files `data/data/tape*.sgm.gz`.

Time period The corpus contains articles from the *The Financial Times* editions from the year 1993.

The marking up was done semi-automatically by means of PERL scripts and some final hand-editing. The PERL scripts are located in the directory `data/prep/`.

3.2 Structure of the corpus

Each file of this corpus consists of a <div0> containing a number of <div1>s representing different articles from the Financial Times. Each <div1> consists of:

- a <div2>, which contains the articletext. The structure of these articles is HEAD, OPENER, BYLINE and Paragraphs, that can contain optional Figures.
- an <index>, which contains several LIST elements. These lists have a TYPE attribute with value *company, country, industry, types, code* or *people*. See below for more details.
- a <bibl>, as described below.

An article contains the following structure:

1. HEAD - The headline of the article.
2. OPENER - This contains the data of publication and the data of processing by the Financial Times, (encoded as <YYMMDD>FT and <YYMMDD> respectively.
3. BYLINE - This contains the name of the agency or journalist responsible for the article and the place where it was written (coded as a <NAME type=place> element).
4. Paragraphs - a number of <P> elements containing the text of the article divided into paragraphs.
5. BIBL - A bibliographic entry for the article, containing <publisher>, <edition>, <biblscope> (giving the page of the newspaper where this article occurred) and <extent> (giving the approximate size of the article in words) elements.
6. Figures - There may be a final paragraph which contains information about any figures, photographs, graphs, etc which appeared in the printed article, but have been omitted from this corpus.

ID attributes on the articles are the original identifiers given to the articles by the FT.

No sub-paragraph markup has been added, except that omitted figures (in article-final paragraphs) have been coded as e.g.

```
<figure>
<head> Dr Armand Hammer, who led Occidental's diversification </head>
<figdesc>Photograph Omitted</figdesc>
</figure>
```

3.2.1 Structure of the index

Each index consists of a number of lists, that are part of the bibliographic information attached to each article. They consist of a list of ITEMs which depend on the list type as follows:

types - ITEMs are the type of news, covered by the contents of the article.

industry - ITEMs are types of industry connected to the subject matter of the article.

country - ITEMs are names of geographic areas connected to the subject matter of the article.

company - ITEMs are the names of companies/organisations which are referred to prominently in the article.

people - ITEMs are the names of people referred to in the article.

code - ITEMs are the identifying codes that refer to the items in the list of industry.

The following is an example of one article:

```
<div1 type=article id=id00EGODBAH7FT>
<div2 type=articletext>
<head>
Unesco aims at resources shift </head>
<opener>
Publication <date>931027FT</date>
Processed by FT <date>940714</date>
</opener>
<byline>By REUTER
<name type=place>Paris</name></byline>
<p>
Unesco Director-General Federico Mayor yesterday announced plans to move
resources from staff to programmes, Reuter reports from Paris. About dollars
120m of the proposed dollars 455m budget would go to field units, raising
their share of operational funds from 38.2 to 45 per cent, he told the
organisation's general conference.
</p>
<p>
Basic education, Unesco's top priority, was to get nearly 39 per cent of the
budget (dollars 1m more than in the previous two-year period). Science would
get 22.5 per cent, with culture winning 17.2 per cent, communication 11.2
per cent and social science 10.3 per cent.
</p>
</div2>
<index>
<list type=country>
<item> FR  France, EC </item>
</list>
<list type=industry>
<item> P9611 Administration of General Economic Programs </item>
</list>
<list type=types>
<item> NEWS  General News </item>
</list>
<list type=code>
<item> P9611 </item>
</list>
</index>
<bibl>
<publisher>The Financial Times</publisher>
<edition>International</edition>
<biblScope>Page 8</biblScope>
<extent>128</extent>
</bibl>
</div1>
```

3.3 Structure of the original

The corpus is a concatenation of articles. The format of Financial times full text source data is shown below:

```
..AN.-OOBAOBKAGFFT
..HL.-
910115FT 910115 Occidental writes off Dollars 2bn in post-Hammer shake-up
(398)
..BL.-

   By MARTIN DICKSON
..DL.-
   NEW YORK
..TX.-
JUST FIVE weeks after the death of Dr Armand Hammer, Occidental Petroleum's
......
dividend. The moves will mean a Dollars 2bn fourth-quarter write-off.
..TX.-
The announcement by Mr Ray Irani, the energy group's new chairman, sharply
......
Occidental into one of the US's top 20 corporations by revenue.
..DS.-

The Financial Times
..XP.-
London Page 19 Photograph Dr Armand Hammer, who led Occidental's
diversification (Omitted).
*****************************************************************
```

Each acticle is built up from a series of sections. Each section contains a marker of the form `..XX.-`(where markers are per the following list), indicating the start of a particular type of data section. A line of at least 64 asterisks(***) indicates the end of each article.

..AN.- Accession Number
The twelve character string is unique to this document within the FT PROFILE database.

..HL.- Headline; breaks down into:

- `yymmddFT` indicates the *date of the publication* of the document and its 'source', The Financial Times.
- `yymmdd` is a date first processed by FT PROFILE, and is not significant to the data content.
- `Text..` represents the *headline proper.* The last 'word' in the headline is a number inside parentheses, indicating the *approximate number of words* within the document.

..BL.- Byline (optional)
Provides the *author*.

..DL.- Dateline (optional)
Provides a *place*.

..TX.- Text (recurring)
Each `..TX.-` indicates the start of a new *paragraph*.

..DS.- Data Supplier
Contains the name of the Supplier or Source

..XP.- eXtended Page; contains the following information:

- Edition Name
- Page number
- Omitted data – When this appears it is of the form

  `xxxx text..... (Omitted).`

  - `xxxx` is one word indicating the nature of the omission (Photograph, Graph, Map, etc.)
  - `text` is any caption which appeared in the paper against the item omitted. This text is optional.
  - `(Omitted)` indicates the 'end' of the omission, whether or not any text was included.

3.3.1 Existing markup

The above mentioned markers at the beginning of each section are the only form of markup that the original data contained. These are sufficient to mark-up the basic structure of each document. With a simple perl-programme, these markers have been converted to TEI-conformant markup. The only 'major' restructuring needs to be done on the `..HL.-`-section; the two dates will be put in a separate **date**-marker, and the wordcount in a separate **extent**-marker.

3.3.2 Markup Process

The raw data that came of the tapes had one problem; they looked like this:

```
                                        ************************************
****************************              ..AN.-OODBICOABRFT
                                            ..HL.-

                                              930209FT 930209 International Co
mpany News: SEC suit threatens NY Post deal     (390)
                                               ..BL.-
```

```
                                              By KAREN ZAGOR
                                               ..DL.-

                                                 NEW YORK
                                                  ..TX.-
                                                    THE FATE of
the New York Post appears to be once more in the balance         following news
that the Securities and Exchange Commission has sued the         tabloid's
prospective new publisher, Mr Steven Hoffenberg.                      ..TX.-

                                                                     The SEC
suit, against M
```

In order to fix this the files have been run through two perl scripts: *transform1.perl* and *transform2.perl*. The result of these transformations, which are the original data are stored in the directory `data/original/new/`.

The actual SGML markup was done by running all thes files through perlscripts *convert1.perl*, *convert2.perl*, *convert3.perl* and *convert4.perl*. The TEI-headers have been added automatically by running the files through a little program called *add-headers*, which also counts the size of the files.

All files have been checked for their conformance to the DTD. The DTD is the general newspaper.dtd.

Chapter 4

The French *Le Monde* Corpus

Corpus of articles from the French newspaper *Le Monde*, consisting of two years (1992-1993) of extracts from the *Le Monde*, being articles with category codes ECO, MDE and INI, approximately ten million words.

4.1 The SGML Corpus

The MLCC *Le Monde* corpus consists of the files

- data/fr01A1292.**NN**.sgm where **NN** = 01 ... 30
- data/fr01A1293.**NN**.sgm where **NN** = 01 ... 32

consisting, respectively of articles from 1992 and 1993, in date order.

The structure of each file is described by the newspaper.dtd DTD file, and have been checked as conformant to this DTD.

The global structure of each file is (indentation of SGML -markup is there to show structure, it is not present in the corpus files):

```
<tei.2>
  <teiheader> ... </teiheader>
  <text id="ID" lang=fr>
    <body>
      <div0 type=storylist>
        <div1 type=article n=N id=ARTICLE_ID> ... </div1> *
      </div0>
    </body>
  </text>
</tei>
```

The structure of each article "<div1 type=article" is roughly as follows, full details can be found in DTD/newspaper.dtd :

```
<div1 type=article n=N id=ARTICLE_ID>
  <opener>
    <date>DATE</date>
      DOC= ...
      FAB= ...
      NUM= ...
      REF= ...
      SEC= ...
      TAI= ...
  </opener>
  <index>
    <list type=TYPE><item>...</item>*</list>*
  </index>
  <div2 type=articletext n=1>
    <div3 type=introduction>
      <p>...</p>*
    </div3>
    <opener><p>...</p>*</opener>
    <byline>...</byline>
    <head type=superheading>...</head>
    <head type=headline>...</head>
    <p> ... </p>*
  </div2>
</div1>
```

The corpus consists of 16977 articles.

4.2 Original Data Format

The original files as received from *Le Monde* were two large files:

- 01A1292 Lines: 555073 Words: 4,220,400 Chars: 28244086
- 01A1293 Lines: 577811 Words: 4,685,735 Chars: 31413192

The character code used in these files is ISO-LATIN-1.

In order to make more manageably sized files, these two large files were split at lines of the form

```
<#FIELD NAME = ACC>.*</#FIELD>
```

into the files

```
orig/fr01A12{92,93}.NN
where NN = 01 ... 30 for 92
         = 01 ... 32 for 93
```

The directory orig contains the original texts as delivered from *Le Monde*. They are marked up in a form of nonstandard SGML, with each element being coded as

```
<#FIELD NAME = name>field_value</#FIELD>
                ---- -----------
```

where name and field_value are variables.

4.3 First stage conversion to SGML

The PERL script prep/prep.perl was used first to convert the original markup into valid SGML (described by the DTD prep/lemonde.dtd).

The original markup was divided into a number of different classes:

(1) The body of the articles were marked with the following markup.

FIELD NAME	Nature of markup
TIJ	Headline Converted into <head type=headline>
TIC	Another kind of headline text? Converted into <head type=superheading>
TIO	A subheadline after headline Converted into <head type=subheading>
SIG	Signature - a persons name (or list of names) (upper case: Surname Forename) Converted into <byline>
CHA	A small paragraph of text (occurs immediately after <ACC>?) Converted into <div3 type=introduction>
ORI	PLACE from our correspondent Converted into <opener>
NTE	Some text (A footnote?) Converted into <note>
TEX	The text of the article. Converted into <div2 type=articletext n=N>

(2) Articles have identification information in the following markup.

FIELD NAME	Nature of markup
ACC	number Signals the start of an article. Converted into <div1 type=article n=N id=idnumber>
DAT	date(yymmdd)

	Converted into <date> element inside <opener>
DOC	One of the following: (BCL|BHL|DAR|DRX|FLA|JGB|JPD|LEY|LLY|MHB|MHC|MYR|RIP)
PUM	always = QUO
REF	reference_number(n-nnn-nn)
SEC	section, one of: (ECO|MDE|INI)
TAI	number
FAB	number
NUM	article_number(date-reference)

The DOC, FAB, NUM, REF, SEC, and TAI fields were converted into text inside the <opener> element.

(3) Bibliographic or keyword information in the following (mostly optional) markup:

FIELD NAME	Nature of markup
FR2	A list
PE2	A list
ET2	A list
DOS	A list of codes from (ECO|GEN|EXT|EX|EC)
AUO	A list of person names
CAT	A list of article category types, when not normal text article. Possible values are: BIOGRAPHIE, BULLETIN, CARTE, CHRONOLOGIE, CORRESPONDANCE DESSIN, DOSSIER, ENTRETIEN, GRAPHIQUE, INTEGRAL, MARCHES MINISTRES, NECROLOGIE, OPINION, ORGANIGRAMME, PHOTO PORTRAIT, PUBLICITE, RECAPITULATIF, RECTIF, SERIE SIX CROCHETS, SUPPLEMENT, TABLEAU
COM	A comment (very rare)
ET1	Comma sep list of org names?
FR1	nl sep list of ?
GO2	One of the following: (BALLET|EXPOSITION|FILM|LIVRE,RAPPORT| LIVRE|PUBLICATION|RADIO|RAPPORT|TELEVISION,FEUILLETON| TELEVISION,FILM|TELEVISION)
LIE	Usually a (list of) article_number(s) Cross references?
PE1	A nl separated list of person names (occurring in article)
SU1	A code(||)
IMA	Very rare. Ignored.
NBI	Very rare. Always has value "1". Ignored
GO1	Very rare.
SIP	Very rare. Ignored
SOT	Very rare. Ignored
SU2	Very rare.

The fields in this section were converted into <list> elements inside an <index> element, one per aticle. Thus the original:

```
<#FIELD NAME = ACC>224745</#FIELD>
<#FIELD NAME = DAT>920101</#FIELD>
<#FIELD NAME = DOC>BHL</#FIELD>
<#FIELD NAME = DOS>EXT,ECO</#FIELD>
<#FIELD NAME = ET2>FRANCE,ESPAGNE,CAISSE D'EPARGNE</#FIELD>
<#FIELD NAME = FAB>12311070</#FIELD>
<#FIELD NAME = NUM>920101-2-014-33</#FIELD>
<#FIELD NAME = PE2>LYONNAISE DUMEZ,CAIXA</#FIELD>
<#FIELD NAME = PUM>QUO</#FIELD>
<#FIELD NAME = REF>2-014-33</#FIELD>
<#FIELD NAME = SEC>ECO</#FIELD>
<#FIELD NAME = TAI>31</#FIELD>
<#FIELD NAME = TIJ>Lyonnaise-Dumez resserre ses liens avec la Caixa</#FIELD>
<#FIELD NAME = TEX>NO PHYSICAL FILE</#FIELD>
```

was converted into:

```
<div1 type=article n=1 id=id224745>
<opener>
<date>920101</date>
DOC=BHL
FAB=12311070
NUM=920101-2-014-33
REF=2-014-33
SEC=ECO
TAI=31
</opener>
<index>
<list type=DOS><item>EXT</item><item>ECO</item></list>
<list type=ET2>
<item>FRANCE</item>
<item>ESPAGNE</item>
<item>CAISSE D'EPARGNE</item>
</list>
<list type=PE2>
<item>LYONNAISE DUMEZ</item>
<item>CAIXA</item>
</list>
</index>
<div2 type=articletext n=1>
<head type=headline>Lyonnaise-Dumez resserre ses liens avec la Caixa</head>
```

(a) Fields of the form

```
<#FIELD NAME = TEX>NO PHYSICAL FILE</#FIELD>
```

were ignored.

(b) In the original paragraphs inside <TEX> were marked with a line final "<" character. These have been converted into standard SGML style paragraphs.

4.4 Second stage conversion of markup to valid SGML

After prep/prep.perl had been run, prep/prep2.perl was run to reorder the SGML structure marked up. Elements which had been noted as being of index or bibliographic type were reordered into global <IDENT> and <BIBL> elements, one per article. The resulting corpus documents were conformant to the prep/lemonde.dtd DTD file.

Then prep/prep3.perl was run to convert the SGML markup to be conformant to the dtd/newspaper.dtd DTD.

prep/prep4.perl contains some last minute modifications.

4.5 Hand correction of corpus

After the corpus had been checked to be SGML conformant to the dtd/newspaper.dtd, the corpus was hand checked and some errors in the headlines were corrected, since some spaces had been omitted between words, e.g.

> <#FIELD NAME = TIJ>La préparation du 44 congrèsLa CGT va sensiblement renouveler ses instances dirigeantes</#FIELD>

has been corrected to

> <head type=headline>La préparation du 44 congrès. La CGT va sensiblement renouveler ses instances dirigeantes</head>

We have tried to catch as many of these as possible, but cannot guarantee complete coverage.

Chapter 5

The German *Handelsblatt* Corpus

The MLCC Financial Newspaper Corpus German part, consisting of articles from *Handelsblatt* from the period 02.01.1986 to 15.06.1988.

5.1 Original data

We received a tape from *Handelsblatt*, which we had a 'little' difficulty reading. However after a number of attempts, we discovered that the tape contained a number of files as follows:

File	Size (Mb)	Desc
1.data	1177	Index and data (data starts at character position 24520000(octal))
2.data	2	Adverts
3.data	2	Adverts
4.data	5	Adverts
5.data	3	Adverts
6.data	1	Adverts
7.data	0	List of American States
8.data	493	Data

We decided to extract approximately 30 Mb of text from the start of the text data in the file 1.data and use that as the MLCC subcorpus. The extracted text was split into articles at lines containing the string "HB Nr." and stored as the files orig/hb.NNN where NNN runs from 001 to 210.

The extracted data appears to consist of the editorial content of *Handelsblatt* between 02.01.1986-15.06.1988. Edinburgh still has the complete data and it would be possible to extract more or a different range of articles if desired.

5.2 Conversion to SGML

The original files contained a large amount of non-textual control characters. These have largely been removed. The preparation process was done by a number of PERL scripts which are in the prep directory.

Prep.perl divides text into articles and structures them with SGML markup. It removes non-printing characters and attempts to classify them as paragraph breaks or as pieces of extraneous text, coded as <err> ... </err> elements.

Prep2.perl restructures the SGML index elements and tidies up the SGML markup provided by perp.perl.

Prep3.perl converts null bytes (octal 000) into newlines.

Finally the add-headers program was used to add TEI headers to the data files.

5.3 Notes on SGML structure

(*) Bylines have not been marked up in SGML. They should be. E.g.

> vwd NEW YORK. Die Preise fuer Kupfer und Aluminium, die infolge

should be marked up as

> <byline> vwd NEW YORK. </byline> Die Preise fuer Kupfer und Aluminium,
> die infolge

(*) The corpus does not use iso-latin-1 characters for non-standard ASCII characters, i.e. umlauted vowels and the 'scharfes s'. Instead they are coded as respectively ae, oe, ue and ss. When they appear capitalised, then both letters of the digraph are capitalised. We have not converted these into iso-latin-1 characters ä, ö, ü and ß, since to do so, would require more on-line lexical information than we had readily available.

(*) <err> elements.

As mentioned above, in the process of extracting the textual data from the original melange of text and other data (probably indexing and formatting information), in a few cases, short pieces of text were found which were difficult to classify. These have been marked up inside <err> ... </err> elements.

These appear to fall into a number of different classes, as follows:

(1) Short bits of gibberish, presumably originally part of control information. For example

> (7,3) Mill. DM, darunter 2,5 (1,6) Mill. DM Grenzzonensonderabschreibungen.
> <err>`S$\_$%!G7`</err>
> </p>
> <p>
> Der Jahresgewinn der Mutter fiel vor Koerperschaftsteuern und der

These can be removed without damage to the integrity of the text.

(2) Short pieces of disembodied text. These appear to be often a piece of text which is missing from somewhere else in the text (also marked with an <err> element). Perhaps originally some text which overflowed from another article. For example

```
... nachzukommen.
<err>orp rutschte 1 /</err>
</p>
<p>
sug 47 /, Exxon / auf 50, Phillips / auf 10 und Mobil / auf 28 /. Zu den ...
```

These can sometimes be linked up with another <err> element which is missing some text. This requires human intervention and an understanding of the text.

(3) Marking missing text at the end of a paragraph

For example:

```
... sich in der vorigen Woche in New York auf, wo sie den Glaeubigerbanken die
<err>neue durch den OElpreissturz geschaffene Lage darlegten. Citi</err>
</p>
<p>
Neben dem Dow Jones schlossen auch die uebrigen Boersenindices im Plus ab.
Der ...
```

The previous example probably continues this one, i.e. the above text should probably be:

```
... sich in der vorigen Woche in New York auf, wo sie den Glaeubigerbanken die
neue durch den OElpreissturz geschaffene Lage darlegten. Citicorp rutschte 1 /
<missingtext>
</p>
<p>
Neben dem Dow Jones schlossen auch die uebrigen Boersenindices im Plus ab.
Der ...
```

Due to lack of time to hand check and correct all these cases, they have been left as is. In the first seven files, they have been edited and some have been edited into <break> milestone markers.

Thus the corpus contains <err> and <break> elements. <err> means a piece of text which belongs to one of the categories above. <break> is a milestone tag (empty element) which marks a place where some text is missing, usually at the end of an paragraph.

Chapter 6

The Italian *Il Sole 24 Ore* Corpus

6.1 Description of the corpus

The *Il Sole 24 Ore* newspaper have provided the MLCC project with a corpus of articles from their archives. The following is a short report on the nature of this data.

The corpus described here contains articles from *Il Sole 24 Ore* from the year 1992. The corpus consists of 26 files. The size of the *Il Sole 24 Ore* data is approximately 13.3 megabytes and contains 1.88 million words. The raw data (as received from the Pisa) are in the files `orig/sole*.txt.gz`. The SGML marked-up version are in the files `data/sole*.tei.gz`.

The *Il Sole 24 Ore* corpus has been converted to TEI P3 conformant SGML markup. This markup was done semi-automatically by means of a PERL script and post-editing. The PERL script is in the file `prep/conv.perl`.

The DTD used by this corpus is a strict subset of the TEI P3 DTD, the subset used is defined in the file `lib/italian.dtd` as an complete DTD.

6.2 Structure of the corpus

Each corpus consists of a number of <DIV1>s which represent newspaper articles. An article consists of the following SGML structure:

OPENER - The initial OPENER element contains a DATELINE element which gives the data of publication of the article (in the form DD-MM-YY).

HEAD - one or more headlines of the article. Headlines are classified into one of the types *rubrica* or *titolazione*. See below for more details.

BYLINE - (optional) gives the name of the author of the article.

OPENER - (optional) This OPENER before the body of the article contains the place in which the article was written.

Paragraphs - Zero or more P elements containing the main text of the article.

Lists - Zero or more LIST elements. These lists have a TYPE attribute with value *area*, *descrittori*, *didascalia*, *evento*, *persone*, *societaoenti*, *tabelle*, or *vedianche*. See below for more details.

TRAILER - Contains a copyright notice on behalf of "Il Sole 24 Ore".

6.2.1 Headlines

Each article has two HEAD elements:

type=rubrica - Gives a broad classification of the topic of the article. It is not clear to us where this text appeared in the newspaper, whether it appeared in the article or at the top of the page.

type=titolazione - The headline of the article.

6.2.2 Lists

These lists, which appear at the end of each article, are part of the bibliographic information attached to each article. They consist of a list of ITEMs which depend on the list type as follows:

area - ITEMs are names of geographic areas connected to the subject matter of the article.

descrittori - ITEMs are key phrases describing the contents of the article.

didascalia - ITEMS are descriptions of photographs or other graphics which appeared in the printed form of the article, but which have been removed from this electronic version. Inside each ITEM is a FIGURE element with attributes n=GRAFICO-$< nn >$, FOTO-$< nn >$, IMMAGINE-$< nn >$ or DISEGNO-$< nn >$, containing a HEAD element giving the caption of the figure.

evento - ITEMs are the names of important events referred to in the articles. This list type is not always specified for an article.

persone - ITEMs are the names of prominent people referred to in the article.

societaoenti - ITEMs are the names of organisations which are referred to prominently in the article.

tabelle - ITEMs are tables which were included in the article. Attributes on the ITEMs are n=TABELLA-$< nn >$ and rend=tabular. Tables are not marked up and white space is significant inside these.

vedianche - ITEMs are a list of references to other articles connected to this one.

6.2.3 Sub-paragraph structure

Paragraphs may contain CORR elements where the MLCC has corrected the text supplied. These elements are of the form

<corr resp=mlcc sic="*original text*">*corrected text*</corr>

Thus the content of the document has been corrected, while the original text has been preserved as the value of an attribute. The major kinds of corrections made was the changing of accented vowels to use ISO-LATIN-1 accented characters rather than apostrophes, (e.g. *Socie'te'* → *Société*).

Throughout we have made a silent correction of word final grave accents (apostrophes following a word final vowel) to the corresponding accented character.

A character histogram shows that the only characters appearing in the corpus that are not 7-bit ascii are the following:

Octal code	char	Octal code	char
300	À	310	È
314	Ì	322	Ò
331	Ù	340	à
350	è	351	é
354	ì	362	ò
363	ó	371	ù

Text may contain the SGML character entities:

&	"&"	ampersand
&lab;	"<"	left angle bracket
&rab;	">"	right angle bracket

6.3 Markup

The original text, as received from Pisa, can be inspected in the `orig` directory. These files were converted to the SGML version by means of the PERL script mentioned above. The output of this script were then hand checked for conformance to the DTD and that we had not made any obvious mistakes in the markup. Finally, with the aid of a native Italian speaker, we marked accented vowels and made a few corrections (marked by CORR elements).

Paragraph boundaries were added, largely automatically, on the basis of the layout of the original texts. In some cases, paragraphs begin with some words in capitals, these should probably be marked up as sub headlines or as additional DATELINE elements inside the body of an article.

Chapter 7

The Spanish *Expansion* Corpus

7.1 Description of the corpus

Size: The size of the *Expansion* data is 72 megabytes and contains around 10 million words. The raw data are in the files orig/expan1**.gz. The SGML marked-up versions are in the files data/expan1**.sgm.gz.

Time period: The corpus contains articles from the *Expansion* editions of 21-10-91 through 24-10-91 and 14-05-94 through 27-12-94.

The structure of each file is described by the dtd/newspaper.dtd DTD file, and have been checked as conformant to this DTD.

7.2 Structure of the SGML corpus

Each file of this corpus consists of a <DIV0> containing a number of <DIV1>s representing different articles from *Expansion*.

The structure of each article "<div1 type=article" is roughly as follows, full details can be found in DTD/newspaper.dtd :

```
<div1 type=article n=ARTICLE_ID>
 <opener><dateline>DATE</dateline></opener>
  <div2 type=articletext>
    <head type=section>...</head>?
    <head type=superheading>...</head>?
    <head type=headline>...</head>
    <byline><name type=person/place>...</name></byline>?
    <div3 type=introduction>
       <p>...</p>*
    </div3>
    <p>...</p>*
    <head type=subheading>...</head>*
```

```
        <div3 type=apoyo>?
           <head type=subheading>...</head>*
           <p>...</p>*
        </div3>
        <div3 type=li>?
           <head type=subheading>...</head>*
           <p>...</p>*
        </div3>
       </div2>
      <bibl>
        <edition>EDITION_NUMBER</edition>
        <biblscope>PAGE_NUMBER</biblscope>
      </bibl>
     </div2>
    </div1>
```

7.3 Original Data Format

The original data as received from Expansion was one big file:

expan1.txt Lines: 1934248 Words: 9,864,348 Chars: 71364608

This file was split at lines of the form

*** BRS DOCUMENT BOUNDARY ***

into the files orig/expan1XX where XX = aa ... bg

The directory orig contains the original texts as delivered from Expansion. Some characters do not conform to ISO-LATIN-1. These characters were replaced by ISO-LATIN-1 conformant characters and can be found in the program replace.perl.

7.4 Conversion to SGML

The PERL scripts prep/prepA.perl, prep/prepB.perl, prep/prepC.perl, prep/clean.perl and prep/clean2.perl were used to convert the original markup into valid SGML (described by the DTD DTD/newspaper.dtd).

What follows is the original structure and the way this was marked up:

..Document-Number	
number	Converted into <div1 type=article n=number>
<NUME>	
NNNN	Edition number
	Converted into <edition> inside <bibl>
<DESD>	

ddmmyyyy	Date
	Converted into <dateline> element inside <opener>
<HAST>	
	Always empty -> ignored
	Converted into <div2 type=articletext>
<NPAG>	
N	Pagenumber
	Converted into <biblscope> inside <bibl>
<PASE>	Always empty -> ignored
<CINT>	Section of the paper in which article appeared
	Converted into <head type=section>
<ANTE>	Superheading
	Converted into <head type=superheading>
<TITU>	Headline
	Converted into <head type=headline>
<ENTR>	
text	a small paragraph of text
	Converted into <div3 type=introduction>
<TEXT>	
text	the article text
<APOY>	Some kind of supporting article
	Converted into <div3 type=apoyo>
<LI>	Some text (quotes, figurecaptions, footnotes ????)
	Converted into <div3 type=li>
<REFI>	Always empty -> ignored
<ANOT>	Always empty -> ignored

Paragraph boundaries were not explicitly marked in the original but could be detected by the sentence-end punctuation followed by whitespaces. If it was the case that there was an end of a paragraph and the next paragraph was one (short) line long, the single line was a subheadline and converted into <head type=subheading>.

In most articles names and places of correspondents or news agencies appeared at the start of <ENTR> or <TEXT>. Names are usually capitalised. They are converted into <name>, <name type=person> or <name type=place> inside <byline>.

7.5 Hand correction of corpus

The corpus was checked to be SGML conformant to the DTD/newspaper.dtd, and some errors were corrected by hand. Some corrections had to be made to the <byline> elements and there were some missing <p> and </p> elements.

Chapter 8

Journal of the European Commission Written Questions Corpus

8.1 The corpus

The text corpus provided to MULTEXT as deliverable 4.2.1 version B consists of extracts from the Journal of the European Commission, Written Questions (1993), henceforth JOCWQ. This corpus contains written questions asked by members of the European Parliament and corresponding answers from the European Commission in 9 parallel versions (languages: Danish, Dutch, English, French, German, Greek, Italian, Portuguese and Spanish). The total size of the corpus is approximately 10.2 million words (~ 1.1 million words per language). This data has been acquired and marked-up as part of the Multilingual Parallel Corpus of the Multilingual Corpora for Cooperation (MLCC) project.

The data was acquired by ISSCO from the Office of Publications of the European Community (OPOCE) and consists of Parliamentary Questions published in 1993.

The deliverable consists of:

1. This document.
2. A TEI conformant and nSGML conformant version of the corpus marked-up to level 1. Some level 2 (sub-paragraph) markup in included where this was easily derived from existing markup.
3. Since some of the transformations applied are non-reversible, the original corpus supplied by EPOCE is included as well, in order to allow errors to be checked.
4. The source of the programs used to automatically markup the corpus files.
5. The source of the TEI P3 DTD.
6. A restricted DTD which describes this corpus. This DTD (which is much smaller than the complete TEI DTD) may be be useful for programs which require the structure of the corpus.

The structure of the corpus is simple. There is a single directory containing 360 (= 40 x 9) files. These correspond to 40 issues of the Written Questions, each in 9 language versions. They are named:

exp.joc<NNN>.93.<LA>.01.tei

where NNN is three digits referring to the issue of the Written Questions and LA is a two letter language code, (da=Danish, de=German, en=English, es=Spanish, fr=French, gr=Greek, it=Italian, nl=Dutch, pt=Portuguese).

8.2 The markup

8.2.1 Definition of Level 1 markup

The JOCWQ corpus has been marked-up with gross structural markup (level 1 markup as defined in the MULTEXT Corpus Encoding Standard (CES) [4]). This kind of markup includes a TEI conformant header, and universal text elements down to the level of the paragraph, e.g. textual divisions (volume,chapter,etc), paragraphs, titles and headings, footnotes, tables. Some CES level 2 markup has been included where this was easily extracted from the original markup, e.g. quoted sections, rendition information and some abbreviations, names and dates.

8.2.2 nSGML Definition

Normalised SGML (nSGML) is a format for SGML marked-up corpora which imposes further restrictions on SGML documents. These restrictions are imposed to (a) improve the readability of corpora and (b) to ease text processing by the MULTEXT tools. A file is in nSGML format if it satisfies the following conditions:

1. Document is a valid SGML document according to some supplied DTD.
2. Document is coded using one of the ISO-LATIN character sets, with embedded character entities where necessary.
3. Reference concrete syntax - processing 8-bit clean in data and attribute values.
4. No capacity/length restrictions.
5. No short refs or tag minimisation.
6. No SUBDOCs.
7. No marked sections.
8. All end-tags present (except for empty elements).
9. All entity references terminated with “;”

10. No SGML elements are broken across multiple lines.

The JOCWQ corpus is in nSGML format as defined above. It also satisfies the invariant that with the exception of #PCDATA inside <HEAD> elements, every line starts with SGML markup. There are no newline characters inside normal <P> elements, ie those which do not contain <FIGURE>s.

8.2.3 The DTD Used

The corpus is conformant to the TEI P3 DTD. In fact, it is conformant to a slightly modified version of the TEI P3 DTD. The only change which we have made is to change the definition of special paragraphs to allowed mixed content, i.e. we redefine

```
<!ENTITY % specialPara '(((%m.chunk), (%component.seq)) | (%paraContent))'>
```

as

```
<!ENTITY % specialPara '(#PCDATA | %m.phrase | %m.inter | %m.chunk)*'>
```

This is necessary to allow quoted paragraphs to be coded as

```
<q>
<p>Paragraph1</p>
<p>Paragraph2</p>
</q>
```

rather than as

```
<q><p>
Paragraph1
</p><p>
Paragraph2
</p></q>
```

In addition to using the standard "TEI.prose" DTD, we also use the "TEI.figures" DTD to allow the markup of tables and figures. Finally, we use a small number of character entities, defined below.

8.2.4 Character set used in JOCWQ corpus

The JOCWQ corpus uses the ISO-LATIN-1 character set with the following exceptions; the Greek files are coded in the ISO-LATIN-7 character set, and the following SGML character entities are used for non ISO-LATIN-1 characters and may occur anywhere in the text of the corpus.

- &dmql; - Double quote mark left
- &dmqr; - Double quote mark right
- &dlqm; - Double low quotation mark
- — - em dash
- – - en dash
- œ - oe ligature (œ)
- ‰ - Permil sign, (per thousand (per million?) 0/00) only occurs once in corpus and was present in original markup.
- ′ - ?, only occurs once in corpus and was present in original markup.
- Ž - Formex G2 character set introduction
-  - Formex G3 character set introduction

The last two character entities occur only where there was an error in the usage of these escape characters, usually within a <SIC> element.

8.2.5 Description of the corpus structure

Each file in the JOCWQ corpus corresponds to one issue of the Journal of the European Commission Written Questions for 1993 in one language. This is represented by a <DIV0> which has a <HEAD> element of type DOCUMENT.BIBLIOGRAPHY and a number of <DIV1>s. The <HEAD> contains various fields and identifiers associated with this document (supplied in original).

Each <DIV1> corresponds to one Written question and an answer to it. Sometimes the final <DIV1> in a document contains corrections to earlier questions. Each <DIV1> has a <HEAD> of type RECORD.BIBLIOGRAPHY containing further EPOCE information about this question and a number of <DIV2> elements.

Each <DIV2> element is either a Written question and its answer (type=wqa) or a correction to an earlier question (type=y). It contains a <HEAD TYPE=OR> which is a textual title for the question, containing an identifier for the question, the name of the person who asked it, and its date. The <DIV2> then contains a number of <DIV3>s.

A <DIV3> contains a number of <DIV4>s of type "Q" (question) or type "R" (reply) which contain the actual text.

Inside a <DIV4> there can be paragraphs <P>, or quoted paragraphs <Q>. The next section describes the possible markup inside paragraphs.

SGML identifiers have been attached to the following elements: (i) The document has an identifier on the <TEXT> element. These are of the form FXA93<*docno*><*langid*>C where <*docno*> is the three digit document identifier and <*langid*> is the two letter language identifier (uppercase). (ii) Each record in the document has an identifier on the associated <DIV1>. These are of the form FXAC93<*docno*><*langid*>C.<*nnnn*>.<*mm*>.00 . <*nnnn*> and <*mm*> are digit strings.

See Table 3 for a formal definition of the structure of the corpus, given in the form of a DTD.

8.2.6 SGML elements inside paragraphs

The following SGML elements may occur inside paragraphs and titles:

- <abbr> - abbreviations such as No., Mrs. etc.
- <num> - sequence numbers for footnotes and footnote references.
- <hi> - used to mark superscript and subscript characters in text.
- <ref> - A reference to a footnote,
  for example, <REF n="note2">(<NUM>2</NUM>)</REF>
  occurs in running text as a reference to the footnote "note2".
- <note> - A footnote.
- <date> - A date. Not all dates are marked-up, only those marked-up in the original.
- <sic> - An error noticed by EPOCE, usually a problem with character set switching, not corrected.
- <corr> - An error noticed and corrected by MLCC, usually a problem with character set switching.
- <name> - The name of a person.
- <rs> - An identifier of a question or answer occurring in text.
- <q> - A section of quoted text.
- <list> - A list of items, normally used for headings for tables.
- <table> - A 2-dimensional table organised into rows and cells.
- <figure> - a more complex type of table.

8.2.7 An example question/answer from the corpus

The following is an example of one complete record from the corpus. Newlines have been inserted to improve readability inside <HEAD> and <P> elements.

<DIV1 TYPE=RECORD ID="FXAC93016ENC.0001.01.00">
<HEAD TYPE=RECORD.BIBLIOGRAPHY>
[RECORD.REF id="FXAC93016ENC/0001/01/00-1" secid="FXAC93016ENC"

type="vert" BIBLEV="s" SEGIND="0" SEGREL="W2"]
[RECORD.COMPL compl="C"]
[RECORD.LA lang="EN"]
[RECORD.MAT vjur="L65" corrid="ORIG" section="C1"]
[RECORD.ID scheme="WQ" year="91" num="1463"]
[RECORD.ID scheme="JO3" year="93" num="01" journal="16"]
[RECORD.BODY role="020" an="PE"]
[RECORD.PART page="1" ordpage="1"]
[RECORD.CLASS class="991E14630000000000" scheme="CLX"]
</HEAD>
<DIV2 TYPE="WQA">
<HEAD TYPE="OR">
WRITTEN QUESTION <ABBR REND="TAIL-SUPER">No</ABBR>
<RS TYPE="WQ">1463/91</RS> by <NAME TYPE=PERSON>
<ABBR REND="TAIL-SUPER">Mrs</ABBR> Brigitte Ernst de la Graete (V)</NAME>
to the Commission of the European Communities
<DATE>(16 July 1991)</DATE>
</HEAD>
<HEAD TYPE="INFO">(93/C 16/01)</HEAD>
<DIV3 TYPE=BODY>
<DIV4 TYPE="Q">
<HEAD>
Subject: Energy cooperation: assessment
</HEAD>
<P>Assessment of the Lomé programmes of action in the ACP countries has brought to light various technical and political problems in respect of energy projects (particularly hydro-electric power stations). It has been shown that certain projects are not feasible because of a lack of funding in the ACP countries concerned, inaccurate economic, environmental and social impact studies or insufficient qualified personnel to supervise and plan the implementation of projects.</P>
<P>Who carried out the assessment and who provided the funding?</P>
<P>Can detailed results be forwarded to Parliament?</P>
<P>What conclusions does the Commission draw from this for future purposes?</P>
</DIV4>
<DIV4 TYPE="R">
<HEAD>
Answer given by <NAME TYPE=PERSON><ABBR REND="TAIL-SUPER">Mr</ABBR>
Marín</NAME> on behalf of the Commission
<DATE>(8 September 1992)</DATE>
</HEAD>
<P>The assessment of energy projects in ACP countries to which the Honourable Member refers was carried out by Sussex Research Associates Ltd of Britain, in tandem with Lahmeyer International of Germany. The assessment was funded under the Article entitled <Q>'evaluation of the results of Community aid and practical follow-up measures'</Q> within the Chapter of the 1986 budget on <Q>'specific measures for cooperation with developing countries'</Q>.</P>
<P>The Commission will forward directly to the Honourable Member and to Parliament's Secretariat a copy of the May 1988 summary report on this evaluation of energy projects in ACP countries.</P>
<P>All assessments seek to make it possible to use the lessons of past experience to improve action in

the future. In this particular case, thanks to the recommendations of the study (which were turned into <Q>'basic principles'</Q> by the meeting of the ACP-EEC Council of Ministers of 6 and 7 May 1991), the departments of the Commission with responsibility for this field now possess guidelines which promise to improve activities in energy-related matters.</P>
</DIV4>
</DIV3>
</DIV2>
</DIV1>

8.2.8 Shortcomings

The main shortcoming with the existing markup is that numbered lists of paragraphs are not marked as such. Also dates and names have not been marked-up inside paragraphs.

8.3 How the corpus was marked up

The corpus came to us marked up in a form of SGML defined in the FORMEX specification [1]. This markup has been converted to TEI conformant SGML [3] as described below. The corpus preparation consisted of the following steps:

1. Automatic Character set conversion.
2. Automatic conversion of FORMEX markup into TEI markup.
3. Hand-editing of TEI markup to ensure conformance with TEI DTD.
4. Automatic writing of TEI headers for the files.

8.3.1 Character set conversion

This section describes how the corpus was transcribed into the ISO-LATIN-1 character set. The Greek texts were already coded in ISO-LATIN-7, so no character set conversion was required for them. In the following, \NNN means the character with numeric value NNN (octal). The corpus used one of two characters as escape characters to indicate that the following character came from a non-standard character set. These two characters are \216 (Formex G2 character set) and \217 (Formex G3 character set).

Table 1 shows how characters in these two additional character sets were translated into ISO-LATIN-1 characters or SGML entities.

Notes

1. Occasionally in the corpus there were <(BLK0)SIC> ... </(BLK0)SIC> elements which included invalid escaped character sequences. These have been replaced by <SIC> ... </SIC> and inside such elements, all invalid occurrances of \216 or \217,

i.e. those which did not match one of the sequences given in table 1, have been replaced by SGML character references e.g. Ž or  .

2. Accents

   There are some occurrances of \216 or \217 which were not in <(BLK0)SIC> ... </(BLK0)SIC> elements but which did not match any of the patterns in table 1. If they matched one of the other FORMEX characters (defined in [2]) and made sense in context, they were replaced by SGML entities. If they looked like errors, e.g. the accent being placed after instead of before a vowel, then they were replaced by

   <CORR RESP=MLCC SIC="Original Text">Replacement Text</CORR>.

   where where "Original Text" was the word which contains the error and "Replacement Text" is what we considered the correct replacement word to be. These changes were made in the hand-correction phase of corpus preparation.

3. Line breaks marked with Ctrl-M Ctrl-J have been replaced by Ctrl-J throughout.

4. Blank lines and line-final whitespace have been removed.

5. Spaces

   The four different varieties (widths) of spaces (i.e. em space, en space, thin space and 4/em space) have all been replaced by space characters. This is a non-reversible operation, but it was considered that the improvement in readability outweighed the loss in information for the purposes of the MULTEXT project.

8.3.2 Conversion of FORMEX SGML into TEI SGML

As indicated before, the corpus came to us marked-up using a number of FORMEX defined SGML DTDs. These have been converted to use the TEI DTD.

Firstly, all SGML processing instructions have been completely removed. This is a non-reversible operation, but it was considered that the improvement in readability outweighed the loss in information for the purposes of the MULTEXT project. With the possible exception of some tables, it appears that little linguistic information has been lost.

Secondly, the FORMEX defined elements have been converted, (mainly one-to-one), to TEI defined elements. The mapping used is defined in Table 2.

A PERL program was developed to do this conversion automatically. This data was first character converted and an SGML DTD was written which described the corpus (the FORMEX DTD's were not made available to us). The data was then converted to a TEI conformant form on the basis of this DTD.

Notes:

1. <(BLK0)NOTES> appears not to code any significant information as it only ever contains a sequence of <NOTE> elements. Thus removal is justified.

2. <(BLK0)TO> and <(BLK0)FOR> are only ever used to markup the word "Commission" (and its translates) and hence it was thought that it did not carry significant

information as the context tells us if the text is TO or FROM the commission. Thus removal is justified.

3. Treatment of bibliographic structure

   Each file contains a bibliographic entry for the file and each question/answer also had a bibliographic header. This information has been converted to text (in a stereotypical form) inside a <HEAD TYPE=DOCUMENT.BIBLIOGRAPHY> or <HEAD TYPE=RECORD.BIBLIOGRAPHY> element inside the relevant <DIV>, rather than into SGML markup. This was done as it did not seem clear how to code this information in TEI. It is still in a form which would allow further processing by later programs.

4. Conversion of SGML identifiers

   SGML identifiers in the original contained '/' characters. As these do not match the SGML declaration we use, all '/' characters inside IDs have been converted to '.' i.e.

   ID="FXAC93006ENC/0001/01/00" ⟹ ID="FXAC93006ENC.0001.01.00"

5. Reformatting and adding paragraph markup

   Each line of text in the original (where it was not structural markup) corresponded to a paragraph of text. Thus each such line has been marked in a <p> ... </p> element. The line breaks have been left as in the original so each <p> element is on a single line.

8.3.3 Hand-editing of TEI markup

After the corpus had been automatically converted, each file was checked for TEI conformance and errors found were corrected by hand. Due to the clean nature of the original texts, relatively few errors were found. Common errors found were:

- Incorrectly placed <Q>, as for example English genitive apostrophes (John's) being taken as quotation marks.
- <P> interfering with <Q>, the major problem with the automatic TEI markup concerned the quotation markup. <Q> elements can occur both inside and outside paragraph and this is what caused problems: sometimes <Q> elements appeared before <P> elements and </Q> elements before </P> elements. Consequently, the editing consisted for the greater part of putting the <Q> and </Q> elements in the right places. When one or more entire paragraphs were quoted, the <Q> elements were put outside the <P> elements; otherwise the <Q> elements were put inside <P> elements at the appropriate place.
- The other type of error that occurred a few times was the need to change <BLKO.EXPL ID= > into <NOTE N= > since the program was not clever enough to catch all of these.
- Errors in the Formex G2 and G3 character sets, as noted in section 4.1 note 2 above.

- Rendition attribute on <HI>, in the original markup, a number of different values were given for the Rendition attribute on <HI> elements. These appeared to be different names for the same thing. Thus, values of "SUP" or "SUF" have been changed to "SUPER" and values of "EXP", "VAL", "INDEX", or "SUB" have been changed to "SUB".
- Change <p REND=tabular N="table "> with a <HEAD> into <p><FIGURE REND=tabular N="table1"><FIGDESC> and put <FIGDESC> after the </HEAD>.
- Put in missing <HEAD TYPE="info"> before text lines of the form (93/.....).
- In all the joc195 files, the list of paragraphs in the last record, together with the <HEAD> were put inside a <LIST> and each pair of paragraphs was put inside an <ITEM>.

8.3.4 Automatic writing of TEI headers

TEI headers were generated by program and filled in with a small amount of document specific data (identifiers, language used, file sizes). These headers are fairly straightforward.

8.3.5 Removal of token internal markup

Originally there were various kinds of token-internal markup occurring in the corpus, e.g.

S<HI REND="SUPER">r.</HI>

as an abbreviation for 'Signor'. These caused problems for segmenter programs, so it was decided to change all token-internal markup to a form which respects token boundaries. For example, the above example of token-internal markup has been replaced by <ABBR> elements with new REND attribute values as follows:

TAIL-SUPER: All but the first character of the token is in superscript characters. e.g.

S<HI REND="SUPER">r.</HI> abbrev. for Signor
→
<ABBR REND=TAIL-SUPER>Sr.</ABBR>

TAIL-SUPER: Token consists of either digits or roman numerals followed by a superscript suffix. e.g.

XIV<HI REND="SUPER">ème</HI>
→
<ABBR REND="TAIL-SUPER">XIVème</ABBR>

TAIL-SUPER: Representation of exponential numbers. e.g. there were a small number of cases of

10<HI REND="SUPER">9</HI> meaning 1,000,000,000

We were not sure how best to represent this last case, presently they are coded as:

10<ABBR REND="TAIL-SUPER">e9</ABBR>

TAIL-SUB: The last character of the token is subscript. e.g.

CO<HI REND="SUB">2</HI> Chemical compound
→
<ABBR REND=TAIL-SUB>CO2</ABBR>

Bibliography

[1] “Formex - Formalised exchange of electronic publications”, edited by C. Guittet, Office for official publications of the European communities, ’New technologies - project management’ department, Luxembourg, 1985.

[2] “Formex V.2/rev - Character sets specifications”, OPOCE, August 1993.

[3] “Guidelines for electronic text encoding and exchange (TEI P3)”, edited by C.M.Sperberg-McQueen and Lou Burnard, ACL, Chicago/Oxford, 1994.

[4] “Specification for the corpus encoding style”, MULTEXT Deliverable 1.5.1, Nancy Ide and Jean Véronis, September, 1994.

Table 1: Character conversions

Original character sequence		ISO-LATIN-1/SGML Replacement	

Character set FORMEX G2

\216)	⟹	‘	Single quote mark left
\2169	⟹	’	Single quote mark right
\2160	⟹	\260	Degree sign
\2161	⟹	\261	± Plus/minus sign
\2163	⟹	\263	^{3} superscript 3
\2164	⟹	\327	× multiplication sign
\2167	⟹	\267	Middle dot
\216#	⟹	\243	£ Pound sign
\216$	⟹	$	Dollar sign
\216*	⟹	&dqml;	Double quote mark left
\216:	⟹	&dqmr;	Double quote mark right
\216’	⟹	\247	§ Section sign

GRAVE ACCENTS

\216Aa	⟹	\340	à
\216Ae	⟹	\350	è
\216Ai	⟹	\354	ì
\216Ao	⟹	\362	ò
\216Au	⟹	\371	ù
\216AA	⟹	\300	À
\216AE	⟹	\310	È
\216AI	⟹	\314	Ì
\216AO	⟹	\322	Ò
\216AU	⟹	\331	Ù

ACUTE ACCENTS

\216Ba	⟹	\341	á
\216Be	⟹	\351	é
\216Bi	⟹	\355	í
\216Bo	⟹	\363	ó
\216Bu	⟹	\372	ú
\216By	⟹	\375	ý
\216BA	⟹	\301	Á
\216BE	⟹	\311	É
\216BI	⟹	\315	Í
\216BO	⟹	\323	Ó
\216BU	⟹	\332	Ú
\216BY	⟹	\335	Ý

CIRCUMFLEX ACCENTS

\216Ca	⟹	\342	â
\216Ce	⟹	\352	ê
\216Ci	⟹	\356	î
\216Co	⟹	\364	ô
\216Cu	⟹	\373	û
\216CA	⟹	\302	Â
\216CE	⟹	\312	Ê
\216CI	⟹	\316	Î
\216CO	⟹	\324	Ô
\216CU	⟹	\333	Û

TILDE ACCENTS

\216Da	⟹	\343	ã
\216Do	⟹	\365	õ
\216Dn	⟹	\361	ñ
\216DA	⟹	\303	Ã
\216DO	⟹	\325	Õ
\216DN	⟹	\321	Ñ

UMLAUTS and DIERESIS

\216Ha	⟹	\344	ä
\216He	⟹	\353	ë
\216Hi	⟹	\357	ï
\216Ho	⟹	\366	ö
\216Hu	⟹	\374	ü
\216HA	⟹	\304	Ä
\216HE	⟹	\313	Ë
\216HI	⟹	\317	Ï
\216HO	⟹	\326	Ö
\216HU	⟹	\334	Ü

Circle above (Danish only)

\216Ja	⟹	\345	å
\216JA	⟹	\305	Å

OTHERS

\216Kc	⟹	\347	ç
\216{	⟹	\337	ß
\216z	⟹	œ	œ
\216a	⟹	\306	Æ
\216q	⟹	\346	æ
\216+	⟹	\253	Angle quote mark left
\216;	⟹	\273	Angle quote mark right

\216y	⟹	\370	ø
\216i	⟹	\330	Ø
\216Q	⟹	\271	Superscript 1
\216?	⟹	\277	¿
\216!	⟹	\241	¡
\216k	⟹	\272	up-o
\216c	⟹	\252	up-a

Character set FORMEX G3

\217+	⟹	-	Hyphen	
\217)	⟹	—	em dash	
\217*	⟹	–	en dash	
\217˝	⟹	&dlqm;	Double low quotation mark	
\217$	⟹		4/em space	(following 4 replaced
\217"	⟹		en space	by spaces)
\217!	⟹		em space	
\217'	⟹		thin space	

Table 2: Conversion of SGML element structure

FORMEX Markup		TEI Markup
<(BLK0)ABREV>	⟹	<ABBR>
<(BLK0)SEQN>	⟹	<NUM>
<(BLK0)INDIC TYPE= >	⟹	<HI REND= >
<(BLK0)EXPL IDREF= >	⟹	<REF N= >
<(BLK0)EXPL ID= >	⟹	<NOTE N= >
<(BLK0)NOTES>	⟹	nothing
<(BLK0)TI>	⟹	<HEAD>
<(BLK0)DATE STD=NO>	⟹	<DATE>
<(BLK0)SIC>	⟹	<SIC>
<(BLK0)QT>	⟹	<Q>
<(BLK0)PERSON>	⟹	<NAME TYPE=PERSON>
<(SUP)SUP>	⟹	<DIV0 TYPE=DOCUMENT><HEAD>
<(RECORD)RECORD>	⟹	<DIV1><HEAD>
<(BLK0)BLK0>	⟹	<DIV2>
<(BLK0)BLK1>	⟹	<DIV3 TYPE=BODY>
<(BLK0)BLK2>	⟹	<DIV4>
<(BLK0)ID>	⟹	<RS>
<(BLK0)TO>	⟹	nothing
<(BLK0)FOR>	⟹	nothing
<(BLK0)TBL ID= >	⟹	<P> <FIGURE REND=TABULAR N= >
<(BLK0)CORPUS>	⟹	<TABLE>
<(BLK0)ROW>	⟹	<ROW>
<(BLK0)COL ID= >	⟹	<CELL N= >
<(BLK0)BLKCOL>	⟹	<LIST TYPE=TABLE.COLUMN.HEADINGS>
<(BLK0)COL1>	⟹	<ITEM>
<(BLK0)COL2>	⟹	<ITEM> inside <LIST>
<(BLK0)NOTCOL>	⟹	<HEAD TYPE=COLUMN.NOTE>
<(BLK0)BLKROW>	⟹	<LIST TYPE=TABLE.ROW.HEADINGS>
<(BLK0)ROW1>	⟹	<ITEM>

Chapter 9

European Parliament Debates Corpus

9.1 The Corpus

This second parallel corpus defined for the MLCC collection is the records of Parliamentary sittings published as an Annex to the Official Journal of the European Community under the title *Debates of the European Parliament*. In what follows we will first briefly describe the origin and nature of the data before discussing the physical preparation of the material.

The Parliamentary Debates are a record of what was said by members during the meeting. The texts also contain copies of written input provided to the meeting and other material such as headlines reflecting the structure of the meetings and explanatory comments on the sessions. The original data, from which the translations are ultimately produced, consists of a transcript of the sittings, each member speaking in the language of his choice. This manuscript (known as the "rainbow version") is circulated to all speakers for possible revision before it is sent out for translation. The final version is thus an edited and translated version of the debates, augmented with written documents submitted to the meeting. The data collection, as acquired and prepared for the MLCC corpus, consists of nine parallel versions of the material. The languages are English, Danish, Dutch, French, German, Greek, Italian, Portuguese and Spanish. The texts delivered comprise the Debates of the European Parliament from January 1992 to July 1994.

The language used in this corpus can be characterized as a written version of "formal spoken" language. The transcripts have been cleaned up (for e.g. grammaticality and style) and revised (for e.g. clarity or political motivations). The topics are varied, including discussions of international events and European Union policies as well as day-to-day problems such as traffic conditions in Luxembourg. The debates contain technical discussions as well as personal opinions, and the style ranges from factual and informative to argumentative. The vocabulary is rich and extensive, with an abundance of technical terms, acronyms and proper names as well as common and popular expressions. In a discussion of pollution and traffic problems, for example, the car driver's behavior is characterized with the expression "zoom, zoom, zoom". Though the interactions are generally quite formal and distant, typical of the communication style in such meetings, there are also examples of very personal interactions, e.g. "That is right, I saw you myself". As such, the language used in this collection represents an excellent base for a variety of NLP technology development and evaluation activities.

	IBM-tapes (MOPAS)	IBM-tapes (ASCII)	Bernoulli (MOPAS)	Diskettes (ASCII)
Danish	10 MB		10 MB	10 MB
Dutch	20 MB		10 MB	
English	15 MB		10 MB	
French		25 MB	5 MB	
German			1 MB	40 MB
Greek			1 MB	40 MB
Italian		30 MB	1 MB	
Portuguese	5 MB	20 MB	10 MB	
Spanish	20 MB		10 MB	

Table 9.1: Overview on the data from the European Parliament.

This corpus represents a collection of multilingual parallel data larger than any currently available to the research community in terms of size and number of languages. Given that it is produced on a regular basis and is not encumbered with major privacy and copyright problems, additional material could easily be added for translation studies in e.g. source and target variation per language pair.[1] Such a growing collection can also serve general language studies in important areas such as language coverage and vocabulary growth.

9.2 Data acquisition

Unlike the Written Questions series described in chapter 8, the acquisition and subsequent preparation of the Debates required considerably more time and effort than originally foreseen in the MLCC proposal. The physical acquisition alone proceeded in phases over the entire year of 1994. With considerable help from Mr. Brogard, Director of the European Parliament Printing Services, we were able to locate and obtain a full set of the Debates. The majority of the data was available from in-house backups, albeit on different media. The missing subsets were located and eventually obtained from four different printing companies in Europe. The varied sources of the material meant that the data sets were delivered on different storage media and used different formatting conventions. The mark-up ranged from quite simple ASCII to a highly complex typesetting language called MOPAS. The final set of data consisted of 15 tapes, 2 Bernouilli diskpacks and 100 diskettes; an overview is given in Table 1 in terms of languages amounts and formats.

For exploratory data conversion work, a first sample tape of data was provided in April 1994. The purpose of this sample tape was first to assure that the data could physically be accessed on the media in which it was provided (1/2 inch reel tape). Once the data itself had been extracted, initial data conversion tests could begin. The data delivered on the Bernoulli disk packs also required the acquisition and installation of a Bernouilli tape reader in order to access the data. The full set of data (300 Mb) was delivered in phases throughout the year

[1] There are eleven sittings each year which represent 250-300 pages per sitting/per language. One page contains approximately 1000 words. The collection of additional data in this series should also be considerably easier since, as of July 1995, the data preparation process for Parliament texts has been modernized.

1994, the last delivery of a Bernouilli disk pack and diskettes containing the outstanding data from the independent printers was obtained from the Parliamentary Printing services in December 1994.

9.3 Data Preparation

Before describing the details of this work, we give an overview of the basic steps followed in preparing the data:

1. The data was read and transferred to a Sun Sparc station for processing
2. Background documentation on the MOPAS commands was collected
3. Reference material, i.e. sample printed versions of the texts, were obtained
4. Programs were written (in test and verify cycles) to parse the MOPAS syntax and establish correspondences with the SGML tags to be introduced.
5. Extensive checking mechanisms were developed to verify the results and capture potential problems.
6. A semi-automatic preprocessing phase was introduced to prepare each file for automatic processing.
7. MOPAS commands were replaced with SGML tags by the program.
8. The MOPAS character set was replaced by the standard ISO-Latin-1 character set; Greek characters represent a special case in need of further treatment.
9. Headers were automatically inserted by program.

It is worth noting that the delivery of new data sets required additional loops in each of the steps outlined above. We now turn to a somewhat more detailed description of each of these steps from reading the data to decoding and transforming the codes to SGML tags.

9.3.1 Reading the Data

As described above, the data for the Parliamentary Debates corpus was delivered in phases throughout the year 1994 on three different media. Each set required a different set of resources and programs to read the data and assure that no data was lost or corrupted during this process.

The data from the three storage media were extracted as follows:

- IBM tapes: The data supplied on 1/2 reel tapes were read blockwise via a tape reader on a DEC machine available at the University of Geneva Computing services. The data was then transferred to a Sun Sparc station at ISSCO by FTP. The block headers and some other undecoded part of the file headers (probably encoding time of creations, revision number, etc.) were removed prior to processing of the data.

- Bernouilli disk packs: In order to read the Bernouilli disk packs, the Parliamentary Printing offices lent ISSCO a Bernouilli reader and a copy of the necessary software to communicate with a PC. The software was installed and the data was then copied to the PC and finally to a Sun Sparc station.[2] A Bernouilli disk pack can contain up to 90 Mb of data and is used for PC back-up purposes.
- floppy disks: The data delivered on the floppy disks (100 diskettes) were copied onto a PC and also transferred via a local network to a Sun Sparc station.

It is worth noting that acquiring data on such diverse media requires a well equipped center and access to experienced technical personnel. To this end, contact was established with a technician at the Parliamentary Printing services as well as a local Swiss company, Delta Information Systems SA. This company was able to provide us with extracts from the "Autologic SA MOPAS manual" describing "Files on magnetic tapes" and the section on "Mopas internal codes".

The structure of the data on the varied storage media did not necessarily correspond to the logical file structure produced in the final data set. Sittings were divided across numerous files, organized differently for each data set. Reading and transferring all of this data (300 Mb) thus required a careful manipulation of hundreds of files in order to retain all of the information potentially useful for further processing.

9.3.2 Data Conversion

The majority of the data supplied was coded in the MOPAS format, a powerful procedural type setting language. The conversion of these highly complex formatting codes to descriptive SGML turned out to be a very complex task requiring additional time and resources not foreseen in the initial project proposal. The first step in this task was to acquire documentation on this formatting language and to compare the electronic data with the printed version.

The Parliamentary Printing offices and the Swiss company, Delta Information systems, who work with this formatting language were helpful in providing background information. However, a full and precise description of the macros used for the Debates were no longer available. One reason for this is that the coding schema evolved over time, thus new options were adopted in more recent sittings.

Decoding the typesetting commands in view of inserting SGML tags is more than a simple decryption task. The typesetting commands are of a procedural nature such that no one-to-one correspondence can be established between code and desired tag. Exploratory programs were written to parse the syntax of the MOPAS commands in an attempt to establish correspondences between the codes and the SGML tags to be introduced. This decoding work proceeded in test and verify cycles. Each new data set (and in fact each new sitting) brought in new variants on the coding. Since not all of the codes could be reliably interpreted, extensive checking mechanisms were written to capture problematic cases.

[2] Since this storage medium is not commonly known to the public we had to rely on in-house expertise and some trial and error exercises to extract the data.

After lengthy trial decoding cycles, a good portion of the codes could be analyzed and interpreted in view of transforming them to SGML tags. However, not all of the special sequences of codes could be reliably recognized and thus no simple replacement scheme was possible. Changes from one set of data to another introduced minor modifications and the original data also contained errors. In order to overcome these problems within the time allotted for this project, we decided to introduce a pre-processing phase as preparation to the automatic conversion program.

In practice, the data preparation thus proceeded in a step-wise fashion. The MOPAS files which contained binary control codes representing the mark-up sequences were converted to an ASCII notation. These control codes were then converted to SGML tags and the output was verified by additional checking programs. In case of errors in the SGML output the conversion program was modified for a new test and verify cycle. This work proceeded in iterations up to a point where it became apparent that the return on program refinement could only bring marginal improvements (given the time and resource constraints of the project). At this point, the intermediate files were edited to avoid the errors otherwise produced by the automatic conversion routines.

The following types of errors, due to variations and errors in the original files, were relatively frequent:

- sequences of footnotes were not separated
- the end of a table was not detected
- the beginning of a debate or annex was not detected properly because the words "sitting" or "annex" were misspelled
- for each speaker, the political party and the language used were usually given in brackets. Sometimes an opening or closing bracket was missing.

This edited intermediate version of the files is provided with the deliverables. The majority of the processing (i.e. the automatic conversion of control codes to SGML tags) is realized as a set of automata written in the form of a C-program. The source code comprises 2,500 lines of code and is provided with the data.

As mentioned above, a subset of the data was supplied in a very simple ASCII without formatting. Though it is quite straightforward to automatically convert this to an SGML conformant document, the mark-up is of little value. A simple program could, for example, replace double carriage returns with open and closing paragraph tags. However, the rich information (as derived from the typesetting codes which provide interesting linguistic information e.g. marking headlines and identifying speakers and languages) cannot be reliably reconstructed without quite intensive and intelligent processing. For the ASCII data containing some markup, a set of programs to convert these codes into SGML have been developed in this project and can probably be applied to the rest of this protion of the ASCII corpus without much additional effort. The rather large set of German data wich contains essentially no markup has not been processed.

An overview of all of the data that has been converted is given in Table 9.2 in terms of languages, word counts and formats.

	IBM-tapes (MOPAS 1)	IBM-tapes (ASCII 1)	Bernoulli (MOPAS 2)	Bernoulli 2 (MOPAS 2)	Diskettes (ASCII 2)
Danish	1.7		1.7	1.7	* 1.7
Dutch	3.4		1.7	0.8	
English	2.5	0.8	1.7	2.5	
French		2.0	0.8	1.7	
German			0.2	0.8	* 7.0
Greek			0.2	1.7	* 7.0
Italian		* 5.0	0.2	1.7	
Portuguese	0.8	* 3.4	1.7	1.7	
Spanish	1.7		1.7	1.7	

Table 9.2: Overview on the data from the European Parliament (figures are in million words). Data marked with "*" require additional work.

9.3.3 Some practical considerations

Data preparation of such a sizable collection from the low-level format to a logically structured document is of necessity a step-wise activity including numerous test and verify cycles. This implies that all of the data should be available for sampling and that numerous copies will be stored for intermediate consistency checking. Though physical storage space is no longer an obstacle, it is worth noting that the manipulation of such a large amount of data contained in the numerous files does require investment in an adequate working environment. This includes not only adequate processing power and disk space but also a proper network environment. In the preparation of this data our environment consisted of access to a 1/2 inch tape reader with a network connection, a Bernouilli disk pack reader, a PC and a SUN Sparc station with adequate disk space (2 Gb).

A few additional practical considerations are worth noting in the documentation of this work. We offer these comments as input to future data preparation projects. The issues listed here identify issues that any data preparation project must address but for which there are no simple answers.

- idealized project planning vs. the practical reality
- time investment in development of automatic conversion routines in view of targeted level of detail and accuracy
- reversibility of conversion steps
- automatic reproduction of information (in perhaps different forms)
- choice of semi-automic, fully automatic and human editing

At the outset of the MLCC project an estimate was made of time necessary for the work foreseen. Practical obstacles such as the delivery of data in phases and in different formats as well as the complexity of the formatting language only became apparent in the course of the project and required some readjustment of plans. All of the data is currently in a form useful for NLP activities though some additional work would be desirable as documented in section 9.7.

Investment in data conversion work is a trade-off between time and accuracy. Older typesetting languages are procedural languages and were not intended for automatic conversion to logical structures. There is thus often a break-off point where further development of fully automatic conversion routines is no longer viable or desirable. E.g. for this corpus, the development of a program for fully automatic conversion of the MOPAS codes would probably take longer than hand-editing to correct for mistakes not caught by the program. As discussed above, we opted for semi-automatic conversion, with a hand-editing phase, in view of the time constraints and the estimated feasibility of fully automatic conversion.

In principle, it is desirable to aim for full reversibility, however, in practice this is rarely possible. As documented in the discussion of the preparation of the parallel data (cf. also 8), some non-reversible step were deemed necessary). These are documented and intermediate copies of the data are also provided to record the non-automatic steps. And in any case, as past activities have demonstrated, no corpus is fully error free and thus will always require some hand-editing depending on the required or desired level of precision.

9.4 Data structure and markup

We provide here a summary description of the document structure and coding conventions. The corpus has been marked-up with gross structural markup including an SGML header and universal text elements down to the level of the paragraph, e.g. textual divisions, paragraphs, titles, headings, footnotes and tables. Some markup inside paragraphs has been included where this was easily extracted from the original markup, e.g. information about speakers.

9.4.1 Global Document Structure

The following gives an overview of the document structure:

```
<tei.2>
   <teiheader> ...
   </teiheader>
<text>
    <body>
<debates>
      <cover>
         <title>
         </title>
         <p> ...
         </p>
      </cover>
      <sitting> ...
         <contents>
```

```
                <p> ...
                </p>
              </contents>
              <p> ...
              </p>
             <annex> ...
                <p> ...
                </p>
              </annex>
          </sitting>
       </debates>
        </body>
   </text>
   </tei.2>
```

Tags that can occur anywhere

```
<headline> </headline>
<footnote> </footnote>
<speaker> </speaker>
<pageref> </pageref>          page references in table of contents
<logo> </logo>                logo of the EC office for publications
<table> </table>
```

Each document corresponds to one issue of the Official Journal of the European Communities – Debates of the European Parliament. The contents of the four cover pages are enclosed by <cover> tags. The main document consists of a number of sittings. Each sitting starts with a table of contents. Some sittings include one or more annexes.

9.4.2 nSGML Definition

The data is coded in normalised SGML (nSGML), a format for SGML marked-up corpora which imposes further restrictions on SGML documents. These restrictions are imposed to (a) improve the readability of corpora and (b) to ease text processing by other programs. The files in nSGML format satisfy the following conditions:

1. Document is a valid SGML document according to the supplied DTD.
2. Document is coded using one of the ISO-LATIN character sets, with embedded character entities where necessary.
3. Reference concrete syntax - processing 8-bit clean in data and attribute values.
4. No capacity/length restrictions.
5. No short refs or tag minimization.
6. No SUBDOCs.

7. No marked sections.

8. All end-tags present (except for empty elements).

9. All entity references terminated with ";"

10. No SGML elements are broken across multiple lines.

All of the files with the exception of the Greek data and the files delivered in ASCII are conformant with this definition.

9.4.3 The DTD Used

The document type definition used in preparation and checking of the data is given in appendix A.3. This DTD accounts for all markup specific to this corpus (that was recoverable from the MOPAS codes).

The header elements were automatically generated by program and filled in with a small amount of document specific data. The programs developed for this task as well as the processing was done by LTG. Work on completing this to bring it up to TEI conformance, e.g. renaming "sittings" to "divisions" with type attributes should be quite straightforward.

9.4.4 The Character Set

The Parliamentary Debates corpus uses the ISO-LATIN-1 character set (exception: Greek files will be coded in the ISO-LATIN-7 character set). The following SGML character entities are used for non ISO-LATIN-1 characters and may occur anywhere in the text of the corpus.

```
Entities

&footref;    reference marker for footnote
&dqml;       double quote mark left
&dqmr;       double quote mark right
&linesep;    line separator within tables
&colsep;     column separator within tables
&amp;        ampersand (&)
&lab;        left angle bracket (<)
&hyphen;     hyphen used for word separation at end of line
&vaigu;      v with accent ' over the character
&parsep;     triangle of stars occasionally used for paragraph separation
```

The full list of character conversion codes is supplied as an Appendix to this chapter.

9.5 Sample from the corpus

The following is an extract from an English version of one of the debates files (deb940117-21.en.b):

```
<headline>1. Approval of the Minutes</headline>
<p>
<speaker>President. -</speaker> The Minutes of yesterday's sitting have been
distributed.
</p>
<p>
Are there any comments?
</p>
<p>
<speaker>Oostlander <party>(PPE)</party>. - <language>(NL)</language></speaker>
Madam President, I have a problem with the translation of a resolution, as
recorded in the Minutes, namely the resolution about Bosnia Hercegovina.
I mention in particular paragraph 1(h) of that resolution. I presented the
resolution in my best English, so that English is the original language.
The strange thing is that in paragraph 1(h) my English expression elimination
comes out as dismantling. Consequently as a result of that change in the
terminology of the text handed in, an equivalent of 'dismantling' has been
entered in all other translations. The strange thing is that the translation
into Dutch again contains the correct expression, 'elimination'. This is an
extraordinarily strange state of affairs, so that as a Dutchman I could not
know yesterday on the basis of my own Dutch text that the original English
text had been changed to a different English text. I should be glad if you
would have that investigated because we, as Members of Parliament, must
naturally be able to rely completely on the integrity of the translation
service. I hope to have an answer from you in due course.
</p>
<p>
<speaker>President. -</speaker> We shall check with the original text as you
have described it. Unless there is a more delicate problem, which I do not see,
if it is simply a translation problem, we shall go back to the original text.
</p>
<p>
<speaker>Oostlander <party>(PPE)</party>. - <language>(NL)</language></speaker>
Madam President, it was not a problem of translation as such, but of altering
an original text into another text by the changing of a word in the original
language.
</p>
<p>
<speaker>President. -</speaker> I understood what you said. So I ask my
fellow-Members to tell me whether they are against the idea of going back to
the text as you originally wrote it, before it was presented to the sitting.
There do not seem to be any objections.
</p>
```

As apparent in this example, the language is riched and varied. Interesting additional markup is included such as the coding of the party and language in the speaker fields. However, other additional codes present in the original, as can be seen from the official printed version, were not possible to recover. And unfortunately, in some cases, information has possibly been lost.[3] This situation, though not ideal, reflects similar problems encountered by all commercial products that must accommodate a range of idiosyncratic markup schema currently in use. Given the current lack of standards and only partial convertibility between systems, these texts can also provide test data for robustness issues.

9.6 Data deliverables

In summary we provide an overview of all of the Parliamentary Debates data that is supplied for the MLCC project. The data is classified according to language and format, the amount is given in millions of words. The starred data has not been transformed nor checked for SGML compliance. Note that initial work on converting the markup codes in the Greek data has been undertaken to assure that the program could also handle the MOPAS codes in these files.

As can be seen from the table, the total size of the corpus is approximately 60 million words (~ 5-7 million words per language).

The deliverable of the Parliamentary Debates corpus consists of:

1. The documentation of the acquisition and preparation work as described in this chapter.
2. A full set of the data supplied on tape (cf. annex for a full listing of the files).
3. The data listed in Table 2 (modulo the subset marked with "*") has been SGML marked up to level 1. Some level 2 (sub-paragraph) markup was included where this was easily derived from existing formatting.
4. A copy of the original corpus (as supplied by the Parliamentary Printing Offices of the European Communities) included as well as an intermediate version of the files preprocessed for automatic conversion.
5. The source code of the programs used to automatically markup the corpus files.
6. The Document Type Definition (still in need of minor modifications in order to ensure TEI-conformance)

The directory structure of the corpus corresponds to the processing phases of the data. The file organization reflects the logical organization of the sittings according to when they were held. The files are named as follows.

[3] In the passage cited above, the careful reader will have noticed that the first use of the words *elimination* and *dismantling* as citations are not quoted in the text, though they probably did appear with special typographical markup in the printed version. We did not have access to all of the written material, nor the resources, to verify such potential mistakes.

deb<JJ><MM><D1>-<D2>.<LA>.<TYPE>

where <JJ><MM> are year and month when the debate took place, <D1> is first day and <D2> the last day of the debate. LA is a two letter language code (da=Danish, de=German, en=English, es=Spanish, fr=French, gr=Greek, it=Italian, nl=Dutch, pt=Portuguese), and <TYPE> gives information of the type of the original markup (m=MOPAS from IBM-tapes, b=MOPAS from Bernoulli tapes, a=ASCII from IBM-tapes).

An example of the files corresponding to one session for each of the languages from data initially delivered in MOPAS format on the Bernouilli tapes is as follows:

```
deb940418-22.da.b.gz Session April 18-22, 1994, Danish
deb940418-22.de.b.gz Session April 18-22, 1994, German
deb940418-22.el.b.gz Session April 18-22, 1994, Greek
deb940418-22.en.b.gz Session April 18-22, 1994, English
deb940418-22.es.b.gz Session April 18-22, 1994, Spanish
deb940418-22.fr.b.gz Session April 18-22, 1994, French
deb940418-22.it.b.gz Session April 18-22, 1994, Italian
deb940418-22.nl.b.gz Session April 18-22, 1994, Dutch
deb940418-22.pt.b.gz Session April 18-22, 1994, Portuguese
```

This organization and file naming convention serves two purposes. The date and language codes correspond to the information given for the printed versions.[4] It should be noted that in some cases the sessions, consisting of numerous sittings, may be distributed over more than one file. The addition of a markup code in the last field (minus the "gz" to indicate the file has been compressed) was adopted in view of possible future additional processing. Where consistent errors are found in the codes, they will most likely be specific to one set of data according to the markup source.[5]

9.7 Further Work

As discussed above, the work invested to prepare this corpus required a substantial amount of work not initially foreseen in the project proposal. The corpus as delivered in its current form would satisfy the needs of a large community of researchers in need of more data in their language. However, some additional processing is still necessary for a subset of the corpus and would be desirable for the entire corpus. The work still required to improve the markup of the corpus to a state where it can be considered in a final form is as follows:

- SGML Markup of the rest of the data.

  The sub-directories `diskettes`, `newtapes` and some of `tapesascii` (i.e. tapes 2-6 of the `orig/tapesascii` directory) have not been cleaned up nor SGML marked up.

[4] The written versions are available from official EU publication offices throughout Europe.

[5] In view of future standards on naming conventions, we note that such corpus specific extensions should always be possible.

The `diskettes` data is reasonably clean text. The German part does not have much in the way of formatting which may make it harder to reliably add e.g. paragraph markup. The Greek has been transliterated into Latin characters and should be converted to the ISO-LATIN-7 encoding standard.

The `newtapes` and the `tapesascii` are heavily formatted, but appear to use the same formatting codes as used in tape1 of `tapesascii`. Programs to convert this into SGML have been developed in this project and hopefully can be applied to the rest of this corpus without much additional effort.

- Upgrading of SGML markup to TEI conformant SGML.

  For the whole Debates corpus we would like to improve the SGML markup, mainly to bring it into line with the markup used in the other sub-corpora of the MLCC.

  - Conversion of the markup into TEI style
    e.g. <sitting> to <div1 type=sitting>, etc.
  - Improving the level of markup.
    e.g. better markup of headings, lists, citations, etc.

- Greek data

  The Greek parts of the corpus (i.e. the subcorpora in the `bernou` and `diskettes` directories) need to be checked thoroughly by a native speaker for correctness. The Greek data in `bernou` uses two character codes (Latin and Greek). Currently these files are coded in ISO-LATIN-1 with a number of ad-hoc conventions for code switching and representing non- ISO-LATIN-1 characters. Work is still needed to standardize these codes.

- Hand-editing

  All of the data could benefit from some additional hand-editing. Spot checks have been carried out to evaluate the current state but no systematic correction has been undertaken. Errors such as potentially incorrectly broken paragraphs (e.g. paragraphs ending without any punctuation) could be automatically detected and then verified or corrected by hand-editing.

Assuming a well-equipped center with experience in such data preparation work, we estimate four person/months over an elapse time of two months. Verification and hand-editing of the data would require personnel with knowledge of the nine languages.

Appendix: MOPAS to ISO character conversion

Original character sequence		ISO-LATIN-1/SGML Replacement	
GRAVE ACCENTS			
a\3	⟹	\224	à
e\3	⟹	\232	è
i\3	⟹	\236	ì
o\3	⟹	\342	ò
u\3	⟹	\249	ù
A\3	⟹	\192	À
E\3	⟹	\200	È
I\3	⟹	\204	Ì
O\3	⟹	\210	Ò
U\3	⟹	\217	Ù
ACUTE ACCENTS			
a\2	⟹	\225	á
e\2	⟹	\233	é
i\2	⟹	\237	í
o\2	⟹	\243	ó
u\2	⟹	\250	ú
A\2	⟹	\193	Á
E\2	⟹	\201	É
I\2	⟹	\205	Í
O\2	⟹	\211	Ó
U\2	⟹	\218	Ú
CIRCUMFLEX ACCENTS			
a\9	⟹	\226	â
e\9	⟹	\234	ê
i\9	⟹	\238	î
o\9	⟹	\244	ô
u\9	⟹	\251	û
A\9	⟹	\194	Â
E\9	⟹	\202	Ê
I\9	⟹	\206	Î
O\9	⟹	\212	Ô
U\9	⟹	\219	Û
TILDE ACCENTS			
a\11	⟹	\227	ã
o\11	⟹	\245	õ

n\11	⟹	\241	ñ
A\11	⟹	\195	Ã
O\11	⟹	\213	Õ
N\11	⟹	\209	Ñ

UMLAUTS and DIERESIS

\163	⟹	\228	ä
\171	⟹	\235	ë
\128\7	⟹	\239	ï
\181	⟹	\246	ö
\187	⟹	\252	ü
\162	⟹	\196	Ä
\170	⟹	\203	Ë
\180	⟹	\214	Ö
\186	⟹	\220	Ü

Circle above

\155	⟹	\229	å
\154	⟹	\197	Å

OTHERS

\159	⟹	\231	ç
\158	⟹	\231	Ç
\135	⟹	\223	ß
\153	⟹	œ	œ
\150	⟹	\198	Æ
\151	⟹	\230	æ
\146	⟹	\171	Angle quote mark left
\147	⟹	\187	Angle quote mark right
\157	⟹	\248	ø
\156	⟹	\216	Ø
\138	⟹	\191	¿
\139	⟹	\161	¡
\199	⟹	-	Hyphen
\131	⟹	—	em dash
\145	⟹	–	en dash
\127	⟹		M-space
\126	⟹		N-space
\94	⟹		thin space
\130	⟹	\176	Degree sign
\148	⟹	\177	Plus/minus sign
\144	⟹	\183	Middle dot
\137	⟹	\163	Pound sign
\189	⟹	&dqml;	Double quote mark left
\190	⟹	&dqmr;	Double quote mark right

Appendix A

SGML DTDs

This appendix contains the full source of the SGML -DTDs used by these corpora.

A.1 TEI

The TEI DTD that we use is the P3 TEI DTD modified as follows. We redefine the contents of special paragraphs (e.g. <q>) to allow <p> and #PCDATA inside them, i.e. we declare the entity specialPara to be:

```
<!ENTITY % specialPara '(#PCDATA | %m.phrase | %m.inter |
                         %m.chunk)*'                          >
```

instead of by:

```
<!ENTITY % specialPara '(((%m.chunk), (%component.seq)) |
                         (%paraContent))'                     >
```

A.2 Newspaper

This DTD is used for the newspaper corpora (with the exception of the Italian). It is largely a subset of the TEI P3 with the exception of the <INDEX> element, which we use to keep non-printed information about the articles. As suggested by Nancy Ide, it would be preferable to use the TEI <KEYWORDS> element for this purpose.

```
<!ENTITY % txtchunks "(#PCDATA | corr | list | figure )*" >

<!-- The Corpus Header                        -->
<!-- TEI HEADER and PACKAGING                 -->

<!ELEMENT TEI.2  - - (teiheader, text) >
<!ELEMENT TEXT   - - (body) >
<!ATTLIST TEXT       ID         ID      #REQUIRED
                     LANG       IDREF   #REQUIRED>
<!ELEMENT BODY   - - (div0*) >

<!ELEMENT TEIHEADER - - (filedesc, encodingdesc, profiledesc, revisiondesc) >

<!ELEMENT FILEDESC - - (titlestmt, editionstmt, extent,
                        publicationstmt, sourcedesc)>

<!-- TITLESTMT - Title of file and associated information -->
<!ELEMENT TITLESTMT       - - (title)>
<!ELEMENT TITLE           - - (#PCDATA)>
<!ELEMENT EDITIONSTMT     - - (edition)>
<!ELEMENT EDITION         - - (#PCDATA)>
<!ELEMENT EXTENT          - - (#PCDATA)>
<!ELEMENT PUBLICATIONSTMT - - (p*)>
<!ELEMENT SOURCEDESC      - - (p*)>

<!-- ENCODINGDESC - description of markup used in this file -->
<!ELEMENT ENCODINGDESC  - - (projectdesc, editorialdecl)>
<!ELEMENT PROJECTDESC   - - (p*)>
<!ELEMENT EDITORIALDECL - - (p*)>

<!-- PROFILEDESC - what languages are used in this document -->
<!ELEMENT PROFILEDESC - - (langusage)>
<!ELEMENT LANGUSAGE   - - (language)>
<!ELEMENT LANGUAGE    - - (#PCDATA)>
<!ATTLIST LANGUAGE        ID    ID      #REQUIRED>

<!-- REVISIONDESC - history of changes made to this file -->
<!ELEMENT REVISIONDESC - - (change+)>
<!ELEMENT CHANGE - - (date, respstmt+, item) >
<!ELEMENT RESPSTMT - - (name,resp) >
<!ELEMENT RESP - - (#PCDATA)>
```

```
<!-- THE DATA -->

<!-- DIV0 a document -->

<!ELEMENT DIV0          - - (div1*) >
<!ATTLIST DIV0  TYPE    (storylist)     storylist
                ORG     (composite)     composite >

<!-- DIV1 an article in the document -->

<!ELEMENT DIV1          - - (head* & opener* & index* & bibl*& div2*) >
<!ATTLIST DIV1
                TYPE    (article)       article
                N       CDATA           #IMPLIED
                ID      ID              #IMPLIED>

<!-- HEAD article headline -->
<!ELEMENT HEAD          - - (%txtchunks) >
<!ATTLIST HEAD
                type    (headline|subheading|rubrica|titolazione|
                         superheading|section|normal)   normal >

<!-- OPENER publication date -->
<!ELEMENT OPENER        - - (#PCDATA|date|dateline|p)* >

<!-- INDEX several lists -->

<!ELEMENT INDEX         - - (list|com)* >
<!ATTLIST INDEX
        TYPE    CDATA   #IMPLIED >

<!-- COM a comment -->
<!ELEMENT COM           - - (#PCDATA) >

<!-- LIST of things -->
<!ELEMENT LIST          - - (item*) >
<!ATTLIST LIST
                type    (descriptor|trade|country|company|personname|
                         area|descrittori|didascalia|evento|
                         persone|societaoenti|tabelle|vedianche|
                         normal|industry|types|code|people|
                         ET2|PE2|FR2|AUO|ET1|FR1|PE1|
                         DOS|GO1|GO2|LIE|SU1|SU2)        NORMAL >

<!-- TEXT -->
<!ELEMENT DIV2          - - (head|p|byline|opener|trailer|div3|note)* >
<!ATTLIST DIV2
                N       CDATA           #IMPLIED
                TYPE    (articletext)   articletext >
```

```
<!ELEMENT DIV3          - - (head|p|byline)* >
<!ATTLIST DIV3
                N       CDATA           #IMPLIED
                TYPE    (introduction|apoyo|li) introduction >

<!ELEMENT NOTE          - - (p)* >
<!ATTLIST NOTE
                N       CDATA           #IMPLIED
                TYPE    (introduction)  introduction >

<!-- Paragraphs -->
<!ELEMENT P             - - (%txtchunks) >
<!ATTLIST P
                REND    (TABULAR|NORMAL)        NORMAL
                N       CDATA                   #IMPLIED >
<!-- Type=NORMAL  Used to encode a normal paragraph of running text -->
<!-- Type=TABULAR Used to encode a certain kind of simple table     -->

<!-- CORR Character errors noted and corrected by MLCC -->
<!ELEMENT CORR          - - (#PCDATA) >
<!ATTLIST CORR
                RESP    (MLCC)  MLCC
                SIC     CDATA   #REQUIRED -- original text-->

<!-- BYLINE author -->
<!ELEMENT BYLINE        - - (#PCDATA|name)* >

<!-- DATELINE -->
<!ELEMENT DATELINE      - - (#PCDATA) >

<!-- DATE a date -->
<!ELEMENT DATE          - - (#PCDATA) >


<!-- ITEM An item in a list -->
<!ELEMENT ITEM  - - ((#PCDATA | list | p | figure | corr )*) >
<!ATTLIST ITEM
                ID      ID                      #IMPLIED
                N       CDATA                   #IMPLIED
                REF     CDATA                   #IMPLIED
                REND    (tabular|normal)        normal>

<!-- FIGURE a table with more complex internal structure -->
<!-- FIGURE omitted photographs or tables                -->
<!ELEMENT FIGURE  - - (#PCDATA|head|figdesc|p)* >
<!ATTLIST FIGURE
                REND    (TABULAR)        TABULAR
                N       CDATA            #IMPLIED >

<!-- NAME The name of a person -->
```

```
<!ELEMENT NAME          - - (#PCDATA | corr )* >
<!ATTLIST NAME
                type (person|place) person >


<!-- FIGDESC the body of a FIGURE -->
<!ELEMENT FIGDESC       - - (#PCDATA) >

<!-- TRAILER -->
<!ELEMENT trailer       - - (#PCDATA) >

<!-- BIBL -->
<!ELEMENT BIBL          - - (PUBLISHER|EDITION|BIBLSCOPE|EXTENT)* >
<!ELEMENT PUBLISHER     - - (#PCDATA) >
<!ELEMENT BIBLSCOPE     - - (#PCDATA) >

<!-- Character entities -->
<!ENTITY amp  "&"         -- ampersand           -->
<!ENTITY lab  "&amp;lab;" -- left  angle bracket -->
<!ENTITY rab  "&amp;rab;" -- right angle bracket -->
<!ENTITY dia  "&amp;dia;" -- diamond             -->

<!-- end of file -->
```

A.3 Debates

```
<!-- SGML-Header for the Debates of the  -->
<!-- European Parliament in MOPAS-format -->

<!ELEMENT debates  - - (cover*,sitting+)
                                          +(headline|footnote|footref|
                                            pageref|logo|table|speaker|
                                            party|language)>
<!ELEMENT cover    - - (title?,p*)>
<!ELEMENT sitting  - - (id?,contents?,p+,annex*)>
<!ELEMENT id       - - (#PCDATA)>
<!ELEMENT contents - - (p+)>
<!ELEMENT annex    - - (p+)>
<!ELEMENT (p|title|headline|footnote|
           footref|speaker|party|
           language|pageref|logo|table) - - (#PCDATA)>

<!ENTITY footref "*">
<!ENTITY dqmr "’’">
<!ENTITY dqml "’’">
<!ENTITY linesep "linesep">
<!ENTITY colsep "colsep">
<!ENTITY amp "&">
<!ENTITY parsep "***********">
<!ENTITY hyphen "-">
<!ENTITY vaigu "v’">
<!ENTITY lab "<">
<!ENTITY oelig "oe">
```

A.4 Formex6 DTD for JOCWQ corpus

The following DTD is a subset of the TEI P3 which describes the JOCWQ corpus. It may be of interest as a more precise definition of the JOCWQ markup. In order to use this DTD on the JOCWQ corpus, it is necessary to use the supplied command 'newtei' which uses a non-standard sgmls catalog file to redefine the TEI P3 DTD to this formex6 DTD.

```
<!-- Model DTD for the FORMEX encoded JOC corpus  Version 6     -->
<!-- DTD for MLCC JOC WQ corpus                                 -->
<!-- this is a subset of the TEI P3 DTD                         -->
<!-- includes a definition of the TEI header subset that we use -->

<!ENTITY % tichunks  "(#PCDATA | abbr | q | sic | ref | rs | name |
                       hi | date | corr)*" >
<!ENTITY % txtchunks "(#PCDATA | abbr | q | sic | ref | note |
                       hi | corr | list | table | figure )*" >

<!-- TEI HEADER and PACKAGING -->

<!ELEMENT TEI.2  - - (teiheader, text) >
<!ELEMENT TEXT   - - (body) >
<!ATTLIST TEXT       ID         ID      #REQUIRED
                     LANG       IDREF   #REQUIRED>
<!ELEMENT BODY   - - (div0) >

<!ELEMENT TEIHEADER - - (filedesc, encodingdesc, profiledesc, revisiondesc) >

<!ELEMENT FILEDESC - - (titlestmt, editionstmt, extent,
                        publicationstmt, sourcedesc)>

<!-- TITLESTMT - Title of file and associated information -->
<!ELEMENT TITLESTMT       - - (title)>
<!ELEMENT TITLE           - - (#PCDATA)>
<!ELEMENT EDITIONSTMT     - - (edition)>
<!ELEMENT EDITION         - - (#PCDATA)>
<!ELEMENT EXTENT          - - (#PCDATA)>
<!ELEMENT PUBLICATIONSTMT - - (p*)>
<!ELEMENT SOURCEDESC      - - (p*)>

<!-- ENCODINGDESC - description of markup used in this file -->
<!ELEMENT ENCODINGDESC  - - (projectdesc, editorialdecl)>
<!ELEMENT PROJECTDESC   - - (p*)>
<!ELEMENT EDITORIALDECL - - (p*)>

<!-- PROFILEDESC - what languages are used in this document -->
<!ELEMENT PROFILEDESC - - (langusage)>
<!ELEMENT LANGUSAGE   - - (language)>
<!ELEMENT LANGUAGE    - - (#PCDATA)>
<!ATTLIST LANGUAGE        ID    ID      #REQUIRED>
```

```
<!-- REVISIONDESC - history of changes made to this file -->
<!ELEMENT REVISIONDESC - - (change+)>
<!ELEMENT CHANGE - - (date, respstmt+, item) >
<!ELEMENT RESPSTMT - - (name,resp) >
<!ELEMENT RESP - - (#PCDATA)>

<!-- THE DATA -->

<!-- DIV0 a document -->

<!ELEMENT DIV0          - - (head, div1*) >
<!ATTLIST DIV0
               TYPE    (document)      document >

<!-- DIV1 a record in the document -->

<!ELEMENT DIV1          - - (head, div2*) >
<!ATTLIST DIV1
               TYPE    (record)        record
               ID      ID              #IMPLIED >

<!-- DIV2 a Written Question -->

<!ELEMENT DIV2  - - (head+, div3*) >
<!ATTLIST DIV2
               TYPE    (wqa | y)       wqa >
<!-- WQA is a Written Question Answer -->
<!-- Y   is a correction              -->

<!-- DIV3 -->

<!ELEMENT DIV3  - - (head?, (div4+ | (p | q | list)* ) ) >
<!ATTLIST DIV3
               TYPE    (body)  body >
<!-- Second alternative is for type=Y only -->

<!-- DIV4 -->

<!ELEMENT DIV4  - - (#PCDATA, head, (P | q | #PCDATA )* ) >
<!-- NB The #PCDATA element before the <HEAD> only ever contains a newline-->
<!ATTLIST DIV4
                TYPE   (q|r)   q >
<!-- Type=q is a written question            -->
<!-- Type=r is a reply to a written question -->

<!-- HEAD -->

<!ELEMENT HEAD          - - (%tichunks;) -- title of div-->
<!ATTLIST HEAD
```

```
                TYPE (OR | RECORD.BIBLIOGRAPHY | DOCUMENT.BIBLIOGRAPHY |
                      INFO | COLUMN.NOTE )  OR >

<!-- Type=OR                   Normal heads of <DIV4>s                 -->
<!-- Type=RECORD.BIBLIOGRAPHY   Bibliographic information attached to   -->
<!--                           a DIV1 describing a written question    -->
<!-- Type=DOCUMENT.BIBLIOGRAPHY Bibliographic information attached to   -->
<!--                           a DIV0 describing a document            -->
<!-- Type=INFO                 Attached to a DIV2, gives an identifier -->
<!--                           to a Written Question/Answer            -->
<!-- Type=COLUMN.NOTE          The head in a list of column headings   -->
<!--                           in a table                              -->

<!-- P -->

<!ELEMENT P             - - (%txtchunks;) >
<!ATTLIST P
                REND    (TABULAR|NORMAL)        NORMAL
                N       CDATA                   #IMPLIED >
<!-- Type=NORMAL  Used to encode a normal paragraph of running text -->
<!-- Type=TABULAR Used to encode a certain kind of simple table     -->

<!-- ABBR an abbreviation  -->
<!ELEMENT ABBR  - - (#PCDATA | hi)* >
<!ATTLIST ABBR
                REND    (TAIL-SUPER|TAIL-SUB|NORMAL)    NORMAL >

<!-- DATE a date -->
<!ELEMENT DATE  - - (#PCDATA | abbr | sic )* >

<!-- REF footnote references -->
<!ELEMENT REF   - - (#PCDATA | num | abbr)* >
<!ATTLIST REF
                N CDATA #REQUIRED >

<!-- NOTE footnotes -->
<!ELEMENT NOTE  - - (#PCDATA | num | abbr | q | sic )* >
<!ATTLIST NOTE
                N    CDATA #REQUIRED >

<!-- RS An identifier for questions, only in titles -->
<!ELEMENT RS    - - (#PCDATA) >
<!ATTLIST RS
                TYPE    (WQ)    WQ >

<!-- HI Appears to show character rendition usually type=sup -->
<!-- for superscript characters ?                             -->
<!ELEMENT HI    - - (#PCDATA) >
<!ATTLIST HI
                REND (super|sub) #REQUIRED >
```

```
<!-- Rend=SUPER Superscript -->
<!-- Rend=SUB   Subscript   -->

<!-- NAME The name of a person -->
<!ELEMENT NAME  - - (#PCDATA | abbr | corr | ref | sic )* >
<!ATTLIST NAME
               type (person) person >

<!-- Q quoted material -->
<!--   occurs either (a) inside a paragraph or (b) at paragraph -->
<!--   level containing a number of paragraphs                  -->

<!ELEMENT Q     - - (#PCDATA | p | abbr | ref | sic )* >

<!-- NUM Sequence Number for Footnotes -->
<!ELEMENT NUM   - - (#PCDATA) >

<!-- SIC Errors marked by EPOCE usually in character codes -->
<!ELEMENT SIC   - - (#PCDATA) >

<!-- CORR Character errors noted and corrected by MLCC -->
<!ELEMENT CORR  - - (#PCDATA) >
<!ATTLIST CORR
               RESP    (MLCC)  MLCC
               SIC     CDATA   #REQUIRED -- original text-->

<!-- LIST A list of things -->
<!ELEMENT LIST  - - (head?, item*) >
<!ATTLIST LIST
               TYPE    (NORMAL|TABLE.COLUMN.HEADINGS|
                        TABLE.ROW.HEADINGS)    NORMAL >

<!-- ITEM An item in a list -->
<!ELEMENT ITEM  - - ((#PCDATA | list | p | q | ref )*) >

<!-- TABLE Tabular data -->
<!ELEMENT TABLE  - - (row+) >

<!-- ROW a row of a table -->
<!ELEMENT ROW  - - (cell+) >

<!-- CELL a cell of a table -->
<!ELEMENT CELL  - - ((#PCDATA | ref )*) >
<!ATTLIST CELL
               N       CDATA   #IMPLIED >

<!-- FIGURE a table with more complex internal structure -->
<!ELEMENT FIGURE  - - (head, figdesc) >
<!ATTLIST FIGURE
```

```
                REND    (TABULAR)        TABULAR
                N       CDATA            #REQUIRED >
<!-- FIGDESC the body of a FIGURE -->
<!ELEMENT FIGDESC  - - (%txtchunks) >

<!-- Character entities -->

<!ENTITY amp   "&">
<!ENTITY oelig  "&amp;oelig;"   -- oe as in (French) &oelig;vre     -->
<!ENTITY mdash  "&amp;mdash;"                                         >
<!ENTITY ndash  "&amp;ndash;"                                         >
<!ENTITY dlqm   "&amp;dlqm;"    -- Double low quotation mark        -->
<!ENTITY dqml   "&amp;dqml;"    -- Double quotation mark left       -->
<!ENTITY dqmr   "&amp;dqmr;"    -- Double quotation mark right      -->
<!ENTITY permil "&amp;permil;"  -- permil sign (per 1000) 0/00      -->
<!ENTITY prime  "&amp;prime;"   -- ? occurs once in original markup -->

<!-- end of file -->
```

Appendix B

Licence Agreement forms

The following are the licence agreements which MLCC made with a number of newspapers. They are included here to act as an example of the kind of licence agreement which could be arranged between ELRA and the corpus data providers.

B.1 Example Agreement between Data Provider and the University of Edinburgh

Licence Agreement Form

Letter of Licence for Copyrighted Material

This letter describes the terms of an agreement between **F.T. Business Enterprises Limited**, whose signature appears below, and The University of Edinburgh, acting through the Human Communication Research Centre (HCRC), in which **F.T. Business Enterprises Limited** gives, free of charge, non-exclusive use of a machine-readable version of the copyrighted material described below to The University of Edinburgh, for the purposes of analysis. In particular, HCRC will analyse the materials with a view to investigating their suitability for inclusion in the text corpus being developed by the MLCC project.

The HCRC is a department of The University of Edinburgh which inter alia undertakes research into the automatic analysis of natural language texts. It is a partner in the Multilingual Corpora for Cooperation (MLCC) and Multilingual Text Tools and Corpora (MULTEXT) projects of the European Union.

The MLCC is an activity which collects machine readable text for the purpose of scientific and humanistic research, and distributes it at cost and without royalties.

The MULTEXT project is an activity which seeks to contribute to the development of generally usable software tools to manipulate and analyse text corpora and to create multilingual corpora with structural and linguistic markup.

Accordingly,

1. **F.T. Business Enterprises Limited** agrees to give The University of Edinburgh a machine-readable copy of the copyrighted material described below under (10) and licence to analyse it subject to the constraints specified below under (3) - (9).
2. **F.T. Business Enterprises Limited** hereby warrants that it has full right and authority to grant these presents and that the said copyrighted material is not subject to any third party right or interest.
3. The University of Edinburgh shall be licenced to use the material for the MULTEXT and MLCC projects subject to the provisions of this Agreement.
4. The University of Edinburgh shall be entitled to distribute copies of the material to the other partners in the MULTEXT project, subject to such partner entering into an agreement similar to this one, confining their use to analysis for research purposes and barring further redistribution.
5. This agreement shall come into effect on the 1st November 1994 and shall terminate on the 31st July 1995 unless either party is in breach of the terms of this agreement. In the case of a material breach of this agreement either party may terminate forthwith upon written notice.
6. Both parties exclude all liability of whatsoever nature for direct, consequential or indirect loss, economic loss or loss of profit of whatsoever kind suffered by the other.
7. Upon termination of this agreement, unless another licence is in effect, The University of Edinburgh and MULTEXT partners shall delete all copies of this material.
8. Nothing in this agreement shall be deemed to vest in The University of Edinburgh any legal or beneficial rights in ownership of the material or in any other data supplied by **F.T. Business Enterprises Limited** which at all times shall remain the property of **F.T. Business Enterprises Limited**.
9. Neither party may assign, sub-contract or delegate the rights or obligations under this agreement without the prior written consent of the other, such consent not to be unreasonably withheld in case of an assignment of the full rights and obligations under this agreement by either party to a subsidiary or associated company.
10. The materials consist of the following machine-readable texts: The full text of the editorial content of the Financial Times 1st January 1993 - 31st December 1993.

(Signed) ________________________ Date ______________
F.T. Business Enterprises Limited

(Signed) ________________________ Date ______________
For The University of Edinburgh

List of MULTEXT partners:

- ISSCO, Geneva
- Digital Equipment B.V.
- Fundacion Bosch Grimpera, Universitat Central de Barcelona
- Arbeitsbereich Linguistik, Westfaelische Wilhelms-Universitaet Muenster
- Istituto di Linguistica Computazionale Consiglio Nazionale delle Richerche, University of Pisa
- Sonovision Itep Technologies, France
- SNI/SIETEC Systemtechnik, Muenchen
- Centro de Desarollo Software, Siemens-Nixdorf Sistemas de Informacion, Barcelona
- Department de Filologia, Universitat Autonoma de Barcelona
- Stichting Taaltechnologie, Utrecht Technology
- Rank Xerox Research Centre, France

B.2 Example Agreement between the University of Edinburgh and Data Users

Licence Agreement Form

Licence agreement between Edinburgh and MULTEXT partners for the MLCC Financial Times 1993 corpus

This letter describes the terms of an agreement between **MULTEXT partner**, and The University of Edinburgh, acting through the Human Communication Research Centre (HCRC), in which The University of Edinburgh gives, free of charge, non-exclusive use of a machine-readable version of the copyrighted material described below to **MULTEXT partner**, for the purposes of use within the MULTEXT and MLCC projects.

The HCRC is a department of The University of Edinburgh which inter alia undertakes research into the automatic analysis of natural language texts. It is a partner in the Multilingual Corpora for Cooperation (MLCC) and Multilingual Text Tools and Corpora (MULTEXT) projects of the European Union.

The MLCC is an activity which collects machine readable text for the purpose of scientific and humanistic research, and distributes it at cost and without royalties.

The MULTEXT project is an activity which seeks to contribute to the development of generally usable software tools to manipulate and analyse text corpora and to create multilingual corpora with structural and linguistic markup.

Accordingly,

1. The University of Edinburgh agrees to give **MULTEXT partner** a machine-readable copy of the copyrighted material described below under (10) and licence to analyse it subject to the constraints specified below under (3) - (9).
2. The University of Edinburgh hereby warrants that it has full right and authority from **F.T. Business Enterprises Limited** to grant these presents.
3. **MULTEXT partner** shall be licenced to use the material only for the MULTEXT and MLCC projects subject to the provisions of this Agreement. **MULTEXT partner** further agrees not to redistribute any of the supplied material to any third parties.
4. **MULTEXT partner** agrees to respect the copyright on the supplied material, which is held by **F.T. Business Enterprises Limited**.
5. This agreement shall come into effect on the date of signing and shall terminate on the 31st July 1995 unless either party is in breach of the terms of this agreement. In the case of a material breach of this agreement either party may terminate forthwith upon written notice.
6. Upon termination of this agreement, unless another licence is in effect, **MULTEXT partner** shall delete all copies of this material in their possession.
7. Both parties exclude all liability of whatsoever nature for direct, consequential or indirect loss, economic loss or loss of profit of whatsoever kind suffered by the other.
8. Nothing in this agreement shall be deemed to vest in The University of Edinburgh or in **MULTEXT partner** any legal or beneficial rights in ownership of the material or in any other data supplied by **F.T. Business Enterprises Limited** which at all times shall remain the property of **F.T. Business Enterprises Limited**.
9. Neither party may assign, sub-contract or delegate the rights or obligations under this agreement without the prior written consent of the other, such consent not to be unreasonably withheld in case of an assignment of the full rights and obligations under this agreement by either party to a subsidiary or associated company.
10. The materials consist of the following machine-readable texts: The full text of the editorial content of the Financial Times 1st January 1993 - 31st December 1993 (copyright 1993 **F.T. Business Enterprises Limited**).

(Signed) ______________________ Date ______________
MULTEXT partner

(Signed) ______________________ Date ______________
For The University of Edinburgh

Appendix C

Full list of data on tape(s)

This section gives a complete listing of the files provided as the MLCC deliverable. The final deliverables of the Multilingual Corpora for Cooperation (MLCC) project consist of the following data:

C.1 Toplevel files

00COPYRIGHT Text file containing an overview of the current state of the copyright/distribution rights of the various corpora.

00FILELIST.tex Latex format file containing complete inventory of contents of the MLCC Corpus.

00README Introductory file.

00REPORT.tex This Latex report on the MLCC project.

00TODO A text file itemizing actions which still need to be done. 'Should' be empty by the time you get this data.

MLCC-DUTCH Corpus of articles from the Dutch financial newspaper *Het Financieel Dagblad*, consisting of the editorial content *Het Financieel Dagblad* of alternate weeks in 1992 and 1993. (DTD = newspaper.dtd)

MLCC-ENGLISH Corpus of articles from the British financial newspaper *The Financial Times* editions from the year 1993. (DTD = newspaper.dtd)

MLCC-FRENCH Corpus of articles from the French newspaper *Le Monde*, consisting of two years (1992-1993) of extracts from the *Le Monde*, being articles with category codes ECO, MDE and INI, approximately ten million words. (DTD = newspaper.dtd)

MLCC-GERMAN Corpus of articles from the German financial newspaper *Handelsblatt* from the years 1986-1988. (DTD = newspaper.dtd)

MLCC-ITALIAN Articles from the Italian financial newspaper *Il Sole 24 Ore* covering the years 92/93. (DTD = tei2.dtd)

MLCC-SPANISH Corpus of articles from the Spanish newspaper *Expansion* from 21/10/91 - 24/10/91 and the year 1994. (DTD = newspaper.dtd)

MLCC-JOCWQ Corpus of extracts from the Journal of the European Commission, Written Questions (1993). This corpus contains written questions asked by members of the European Parliament and their corresponding answers from the European Commission. The corpus is in 9 parallel versions (languages: Danish, Dutch, English, French, German, Greek, Italian, Portuguese and Spanish). (DTD = tei2.dtd)

MLCC-DEBATES Corpus of European Parliamentary debates from the year 1992-1994. (DTD = debates.dtd)

BIN Various small programs for accessing the data.

DTD DTDs and other SGML -related data files.

C.2 Sub-corpus structure

Each sub-corpus directory contains the following structure:

00COPYRIGHT File containing details of the copyright holder and distribution rights.

00README File containing overall description of the corpus.

licence Directory containing licence agreements.

data Directory containing the SGML marked-up corpus.

editdecl.txt File describing the markup of the corpus and its conversion from the original corpus.

orig Directory containing the original text of the corpus.

prep Directory containing software and details of the markup/SGML conversion process.

doc.tex LATEX documentation file (a chapter of 00REPORT.tex)

doc.body File containing text body of doc.tex.

C.3 Full file listing

Numbers occuring before file names are file sizes in kilobytes.

MLCC-RELEASE: total 628

4	00COPYRIGHT	1	BIN	1	MLCC-GERMAN
0	00FILELIST.tex	1	DTD	1	MLCC-ITALIAN
4	00README	1	MLCC-DEBATES	1	MLCC-JOCWQ
568	00REPORT.ps	1	MLCC-DUTCH	1	MLCC-SPANISH
41	00REPORT.tex	1	MLCC-ENGLISH		
1	00TODO	1	MLCC-FRENCH		

MLCC-RELEASE/BIN: total 6

1	00README	1	isowc	1	tei
1	check-conform	1	make-filelist	1	textonly

MLCC-RELEASE/DTD: total 25

1	00README	3	debates.dtd	3	newspaper.txt
6	catalog	1	debates.dtd.orig	2	p3-dtd
1	chars.ent	5	newspaper.dtd	3	sgmldecl.tei

MLCC-RELEASE/DTD/p3-dtd: total 335

3	p2ents.dtd	1	teifig2.ent	3	teishd2.dtd
4	p3x.dtd	4	teifron2.dtd	5	teispok2.dtd
2	p3xents.dtd	9	teifs2.dtd	2	teispok2.ent
3	sgmldecl.tei	8	teifsd2.dtd	9	teistr2.dtd
9	tei2.dtd	1	teigen2.dtd	6	teitc2.dtd
5	teiana2.dtd	15	teigis2.ent	2	teitc2.ent
1	teiana2.ent	25	teihdr2.dtd	3	teite2f.dtd
1	teiback2.dtd	3	teikey2.ent	4	teite2n.dtd
2	teicert2.dtd	8	teilink2.dtd	2	teiterm2.dtd
22	teiclas2.ent	3	teilink2.ent	2	teiterm2.ent
21	teiclas2.ent.orig	1	teimix2.dtd	7	teitran2.dtd
30	teicore2.dtd	10	teind2.dtd	1	teitran2.ent
10	teicorp2.dtd	2	teind2.ent	12	teitsd2.dtd
15	teidict2.dtd	6	teinet2.dtd	2	teitsd2a.dtd
5	teidict2.ent	7	teiphys2.dtd	11	teitsd2b.dtd
6	teidram2.dtd	1	teiphys2.ent	3	teivers2.dtd
1	teidram2.ent	2	teipl2.dtd	2	teivers2.ent
3	teifig2.dtd	2	teipros2.dtd	8	teiwsd2.dtd

C.3.1 Parallel Debates Corpus

MLCC-RELEASE/MLCC-DEBATES: total 36

1	00COPYRIGHT	16	doc.body	1	orig
12	00README	1	doc.tex	1	prep
1	00TODO	1	editdecl.txt		
1	data	1	manmod		

MLCC-RELEASE/MLCC-DEBATES/data: total 16

10	bernou	2	tapesascii	4	tapesmopas

MLCC-RELEASE/MLCC-DEBATES/data/bernou: total 64459

640	deb930419-23.da.sgm.gz	712	deb940207-11.de.sgm.gz
624	deb930419-23.en.sgm.gz	768	deb940207-11.el.sgm.gz
672	deb930419-23.es.sgm.gz	632	deb940207-11.en.sgm.gz
672	deb930419-23.nl.sgm.gz	680	deb940207-11.es.sgm.gz
664	deb930419-23.pt.sgm.gz	696	deb940207-11.fr.sgm.gz
648	deb930524-28.da.sgm.gz	664	deb940207-11.it.sgm.gz
632	deb930524-28.en.sgm.gz	680	deb940207-11.nl.sgm.gz
680	deb930524-28.es.sgm.gz	680	deb940207-11.pt.sgm.gz
680	deb930524-28.nl.sgm.gz	136	deb940223-24.da.sgm.gz
680	deb930524-28.pt.sgm.gz	144	deb940223-24.de.sgm.gz
648	deb930621-25.da.sgm.gz	152	deb940223-24.el.sgm.gz
640	deb930621-25.en.sgm.gz	128	deb940223-24.en.sgm.gz
688	deb930621-25.es.sgm.gz	136	deb940223-24.es.sgm.gz
696	deb930621-25.nl.sgm.gz	136	deb940223-24.fr.sgm.gz
688	deb930621-25.pt.sgm.gz	136	deb940223-24.it.sgm.gz
696	deb930712-16.da.sgm.gz	136	deb940223-24.nl.sgm.gz
680	deb930712-16.en.sgm.gz	136	deb940223-24.pt.sgm.gz
736	deb930712-16.es.sgm.gz	91	deb940307-07.da.sgm.gz
736	deb930712-16.nl.sgm.gz	112	deb940307-07.de.sgm.gz
728	deb930712-16.pt.sgm.gz	120	deb940307-07.el.sgm.gz
608	deb930913-17.da.sgm.gz	89	deb940307-07.en.sgm.gz
592	deb930913-17.en.sgm.gz	112	deb940307-07.es.sgm.gz
640	deb930913-17.es.sgm.gz	112	deb940307-07.fr.sgm.gz
640	deb930913-17.nl.sgm.gz	96	deb940307-07.it.sgm.gz
632	deb930913-17.pt.sgm.gz	112	deb940307-07.nl.sgm.gz
128	deb930929-30.da.sgm.gz	95	deb940307-07.pt.sgm.gz
128	deb930929-30.en.sgm.gz	488	deb940308-11.da.sgm.gz
136	deb930929-30.es.sgm.gz	536	deb940308-11.de.sgm.gz
144	deb930929-30.fr.sgm.gz	584	deb940308-11.el.sgm.gz
136	deb930929-30.nl.sgm.gz	480	deb940308-11.en.sgm.gz
136	deb930929-30.pt.sgm.gz	512	deb940308-11.es.sgm.gz
136	deb931013-14.da.sgm.gz	536	deb940308-11.fr.sgm.gz
136	deb931013-14.en.sgm.gz	520	deb940308-11.it.sgm.gz
144	deb931013-14.es.sgm.gz	520	deb940308-11.nl.sgm.gz
152	deb931013-14.fr.sgm.gz	520	deb940308-11.pt.sgm.gz
144	deb931013-14.nl.sgm.gz	128	deb940323-24.da.sgm.gz
144	deb931013-14.pt.sgm.gz	136	deb940323-24.de.sgm.gz
640	deb931025-29.da.sgm.gz	152	deb940323-24.el.sgm.gz
624	deb931025-29.en.sgm.gz	120	deb940323-24.en.sgm.gz
672	deb931025-29.es.sgm.gz	128	deb940323-24.es.sgm.gz
680	deb931025-29.fr.sgm.gz	128	deb940323-24.fr.sgm.gz
664	deb931025-29.nl.sgm.gz	128	deb940323-24.it.sgm.gz
672	deb931025-29.pt.sgm.gz	128	deb940323-24.nl.sgm.gz
600	deb931115-19.da.sgm.gz	128	deb940323-24.pt.sgm.gz
584	deb931115-19.en.sgm.gz	624	deb940418-22.da.sgm.gz
624	deb931115-19.es.sgm.gz	688	deb940418-22.de.sgm.gz
640	deb931115-19.fr.sgm.gz	744	deb940418-22.el.sgm.gz
632	deb931115-19.nl.sgm.gz	608	deb940418-22.en.sgm.gz
616	deb931115-19.pt.sgm.gz	656	deb940418-22.es.sgm.gz
144	deb931201-02.da.sgm.gz	672	deb940418-22.fr.sgm.gz
144	deb931201-02.en.sgm.gz	648	deb940418-22.it.sgm.gz
144	deb931201-02.es.sgm.gz	656	deb940418-22.nl.sgm.gz
144	deb931201-02.fr.sgm.gz	656	deb940418-22.pt.sgm.gz
144	deb931201-02.nl.sgm.gz	616	deb940502-06.da.sgm.gz
144	deb931201-02.pt.sgm.gz	680	deb940502-06.de.sgm.gz
616	deb931213-17.da.sgm.gz	728	deb940502-06.el.sgm.gz
608	deb931213-17.en.sgm.gz	600	deb940502-06.en.sgm.gz
640	deb931213-17.es.sgm.gz	648	deb940502-06.es.sgm.gz
656	deb931213-17.fr.sgm.gz	664	deb940502-06.fr.sgm.gz
656	deb931213-17.nl.sgm.gz	640	deb940502-06.it.sgm.gz
640	deb931213-17.pt.sgm.gz	656	deb940502-06.nl.sgm.gz
608	deb940117-21.da.sgm.gz	648	deb940502-06.pt.sgm.gz
672	deb940117-21.de.sgm.gz	208	deb940719-22.da.sgm.gz
736	deb940117-21.el.sgm.gz	224	deb940719-22.de.sgm.gz
600	deb940117-21.en.sgm.gz	240	deb940719-22.el.sgm.gz
640	deb940117-21.es.sgm.gz	200	deb940719-22.en.sgm.gz
656	deb940117-21.fr.sgm.gz	216	deb940719-22.es.sgm.gz
624	deb940117-21.it.sgm.gz	224	deb940719-22.fr.sgm.gz
648	deb940117-21.nl.sgm.gz	216	deb940719-22.it.sgm.gz
640	deb940117-21.pt.sgm.gz	216	deb940719-22.nl.sgm.gz
648	deb940207-11.da.sgm.gz	216	deb940719-22.pt.sgm.gz

MLCC-RELEASE/MLCC-DEBATES/data/tapesascii: total 14143

344	deb920113-17.en.sgm.gz	648	deb920914-18.fr.sgm.gz
368	deb920113-17.fr.sgm.gz	680	deb921026-30.fr.sgm.gz
544	deb920210-14.en.sgm.gz	672	deb921116-20.fr.sgm.gz
600	deb920210-14.fr.sgm.gz	680	deb921214-18.fr.sgm.gz
72	deb920309-09.en.sgm.gz	696	deb930118-22.fr.sgm.gz
79	deb920309-09.fr.sgm.gz	648	deb930208-12.fr.sgm.gz
448	deb920310-13.en.sgm.gz	112	deb930308-08.fr.sgm.gz
496	deb920310-13.fr.sgm.gz	624	deb930309-12.fr.sgm.gz
544	deb920406-10.en.sgm.gz	648	deb930419-23.fr.sgm.gz
600	deb920406-10.fr.sgm.gz	656	deb930524-28.fr.sgm.gz
632	deb920511-15.fr.sgm.gz	664	deb930621-25.fr.sgm.gz
632	deb920608-12.fr.sgm.gz	704	deb930712-16.fr.sgm.gz
680	deb920706-10.fr.sgm.gz	672	deb930913-17.fr.sgm.gz

MLCC-RELEASE/MLCC-DEBATES/data/tapesmopas: total 31192

384	deb920113-17.es.sgm.gz	704	deb921116-20.es.sgm.gz
384	deb920113-17.nl.sgm.gz	704	deb921116-20.nl.sgm.gz
624	deb920210-14.es.sgm.gz	672	deb921214-18.da.sgm.gz
624	deb920210-14.nl.sgm.gz	656	deb921214-18.en.sgm.gz
592	deb920310-13.es.sgm.gz	712	deb921214-18.es.sgm.gz
592	deb920310-13.nl.sgm.gz	704	deb921214-18.nl.sgm.gz
624	deb920406-10.es.sgm.gz	688	deb930118-22.da.sgm.gz
624	deb920406-10.nl.sgm.gz	720	deb930118-22.es.sgm.gz
608	deb920511-15.en.sgm.gz	720	deb930118-22.nl.sgm.gz
664	deb920511-15.es.sgm.gz	712	deb930118-22.pt.sgm.gz
656	deb920511-15.nl.sgm.gz	664	deb930128-22.en.sgm.gz
608	deb920605-12.en.sgm.gz	640	deb930208-12.da.sgm.gz
656	deb920608-11.nl.sgm.gz	624	deb930208-12.en.sgm.gz
656	deb920608-12.es.sgm.gz	672	deb930208-12.es.sgm.gz
656	deb920706-10.en.sgm.gz	680	deb930208-12.nl.sgm.gz
704	deb920706-10.es.sgm.gz	672	deb930208-12.pt.sgm.gz
704	deb920706-10.nl.sgm.gz	112	deb930308-08.da.sgm.gz
624	deb920914-18.en.sgm.gz	96	deb930308-08.en.sgm.gz
680	deb920914-18.es.sgm.gz	112	deb930308-08.es.sgm.gz
680	deb920914-18.nl.sgm.gz	120	deb930308-08.nl.sgm.gz
672	deb921026-30.da.sgm.gz	120	deb930308-08.pt.sgm.gz
656	deb921026-30.en.sgm.gz	720	deb930309-12.da.sgm.gz
712	deb921026-30.es.sgm.gz	600	deb930309-12.en.sgm.gz
712	deb921026-30.nl.sgm.gz	656	deb930309-12.es.sgm.gz
664	deb921116-20.da.sgm.gz	656	deb930309-12.nl.sgm.gz
648	deb921116-20.en.sgm.gz	648	deb930309-12.pt.sgm.gz

MLCC-RELEASE/MLCC-DEBATES/manmod: total 6

4	bernou	2	tapesmopas

MLCC-RELEASE/MLCC-DEBATES/manmod/bernou: total 90000

864	deb01a94.da.mod.Z	160	deb03c94.en.mod.Z	288	deb07a94.es.mod.Z
928	deb01a94.de.mod.Z	176	deb03c94.es.mod.Z	304	deb07a94.fr.mod.Z
1088	deb01a94.el.mod.Z	184	deb03c94.fr.mod.Z	280	deb07a94.it.mod.Z
816	deb01a94.en.mod.Z	168	deb03c94.it.mod.Z	280	deb07a94.nl.mod.Z
904	deb01a94.es.mod.Z	168	deb03c94.ne.mod.Z	296	deb07a94.pt.mod.Z
936	deb01a94.fr.mod.Z	176	deb03c94.po.mod.Z	856	deb09a93.da.mod.Z
864	deb01a94.it.mod.Z	912	deb04a93.da.mod.Z	808	deb09a93.en.mod.Z
888	deb01a94.ne.mod.Z	856	deb04a93.en.mod.Z	912	deb09a93.es.mod.Z
920	deb01a94.po.mod.Z	944	deb04a93.es.mod.Z	880	deb09a93.ne.mod.Z
904	deb02a94.da.mod.Z	928	deb04a93.ne.mod.Z	920	deb09a93.po.mod.Z
976	deb02a94.de.mod.Z	952	deb04a93.po.mod.Z	168	deb09b93.da.mod.Z
1120	deb02a94.el.mod.Z	880	deb04a94.da.mod.Z	168	deb09b93.en.mod.Z
856	deb02a94.en.mod.Z	976	deb04a94.de.mod.Z	176	deb09b93.es.mod.Z
936	deb02a94.es.mod.Z	1128	deb04a94.el.mod.Z	192	deb09b93.fr.mod.Z
968	deb02a94.fr.mod.Z	832	deb04a94.en.mod.Z	176	deb09b93.ne.mod.Z
880	deb02a94.it.mod.Z	928	deb04a94.es.mod.Z	184	deb09b93.po.mod.Z
920	deb02a94.ne.mod.Z	984	deb04a94.fr.mod.Z	184	deb10a93.da.mod.Z
952	deb02a94.po.mod.Z	888	deb04a94.it.mod.Z	176	deb10a93.en.mod.Z
184	deb02b94.da.mod.Z	896	deb04a94.ne.mod.Z	184	deb10a93.es.mod.Z
192	deb02b94.de.mod.Z	952	deb04a94.po.mod.Z	200	deb10a93.fr.mod.Z
216	deb02b94.el.mod.Z	920	deb05a93.da.mod.Z	184	deb10a93.ne.mod.Z
168	deb02b94.en.mod.Z	864	deb05a93.en.mod.Z	192	deb10a93.po.mod.Z
184	deb02b94.es.mod.Z	968	deb05a93.es.mod.Z	904	deb10b93.da.mod.Z
192	deb02b94.fr.mod.Z	920	deb05a93.ne.mod.Z	840	deb10b93.en.mod.Z
176	deb02b94.it.mod.Z	968	deb05a93.po.mod.Z	944	deb10b93.es.mod.Z
184	deb02b94.ne.mod.Z	864	deb05a94.da.mod.Z	976	deb10b93.fr.mod.Z
184	deb02b94.po.mod.Z	944	deb05a94.de.mod.Z	912	deb10b93.ne.mod.Z
144	deb03a94.da.mod.Z	1080	deb05a94.el.mod.Z	976	deb10b93.po.mod.Z
152	deb03a94.de.mod.Z	816	deb05a94.en.mod.Z	848	deb11a93.da.mod.Z
176	deb03a94.el.mod.Z	896	deb05a94.es.mod.Z	792	deb11a93.en.mod.Z
136	deb03a94.en.mod.Z	944	deb05a94.fr.mod.Z	880	deb11a93.es.mod.Z
152	deb03a94.es.mod.Z	872	deb05a94.it.mod.Z	904	deb11a93.fr.mod.Z
152	deb03a94.fr.mod.Z	880	deb05a94.ne.mod.Z	880	deb11a93.ne.mod.Z
144	deb03a94.it.mod.Z	920	deb05a94.po.mod.Z	912	deb11a93.po.mod.Z
144	deb03a94.ne.mod.Z	920	deb06a93.da.mod.Z	184	deb12a93.da.mod.Z
152	deb03a94.po.mod.Z	856	deb06a93.en.mod.Z	184	deb12a93.en.mod.Z
696	deb03b94.da.mod.Z	960	deb06a93.es.mod.Z	192	deb12a93.es.mod.Z
752	deb03b94.de.mod.Z	944	deb06a93.ne.mod.Z	200	deb12a93.fr.mod.Z
864	deb03b94.el.mod.Z	984	deb06a93.po.mod.Z	192	deb12a93.ne.mod.Z
656	deb03b94.en.mod.Z	976	deb07a93.da.mod.Z	200	deb12a93.po.mod.Z
720	deb03b94.es.mod.Z	920	deb07a93.en.mod.Z	864	deb12b93.da.mod.Z
760	deb03b94.fr.mod.Z	1032	deb07a93.es.mod.Z	824	deb12b93.en.mod.Z
720	deb03b94.it.mod.Z	1016	deb07a93.ne.mod.Z	904	deb12b93.es.mod.Z
720	deb03b94.ne.mod.Z	1048	deb07a93.po.mod.Z	936	deb12b93.fr.mod.Z
744	deb03b94.po.mod.Z	280	deb07a94.da.mod.Z	880	deb12b93.ne.mod.Z
168	deb03c94.da.mod.Z	304	deb07a94.de.mod.Z	912	deb12b93.po.mod.Z
184	deb03c94.de.mod.Z	352	deb07a94.el.mod.Z		
224	deb03c94.el.mod.Z	256	deb07a94.en.mod.Z		

MLCC-RELEASE/MLCC-DEBATES/manmod/tapesmopas: total 41203

480	378.1.mod.Z	816	381.3.mod.Z	744	386.4.mod.Z
768	378.2.mod.Z	864	381.4.mod.Z	128	386.5.mod.Z
720	378.3.mod.Z	824	382.1.mod.Z	848	387.1.mod.Z
6	378.3a.mod.Z	888	382.2.mod.Z	856	387.2.mod.Z
760	378.4.mod.Z	856	382.3.mod.Z	872	387.3.mod.Z
792	378.5.mod.Z	896	382.4.mod.Z	808	387.4.mod.Z
496	379.1.mod.Z	2600	383.1.rainbow.Z	144	387.5.mod.Z
784	379.2.mod.Z	848	384.1.mod.Z	928	388.1.mod.Z
752	379.3.mod.Z	840	384.2.mod.Z	840	388.2.mod.Z
5	379.3a.mod.Z	856	384.3.mod.Z	144	388.3.mod.Z
792	379.4.mod.Z	864	384.4.mod.Z	840	388.4.mod.Z
816	379.5.mod.Z	888	385.1.mod.Z	792	389.1.mod.Z
736	380.1.mod.Z	896	385.2.mod.Z	136	389.2.mod.Z
728	380.2.mod.Z	896	385.3.mod.Z	912	389.3.mod.Z
792	380.3.mod.Z	832	385.4.mod.Z	736	390.1.mod.Z
752	380.4.mod.Z	144	385.5.mod.Z	824	391.1.mod.Z
808	380.5.mod.Z	784	386.1.mod.Z	800	392.1.mod.Z
800	381.1.mod.Z	808	386.2.mod.Z		
856	381.2.mod.Z	808	386.3.mod.Z		

MLCC-RELEASE/MLCC-DEBATES/orig: total 10

4	bernou	1	diskettes	2	newtapes	1	tapesascii	2	tapesmopas

MLCC-RELEASE/MLCC-DEBATES/orig/bernou: total 100213

13	00README	232	deb03c94.elz	304	deb07a94.enz
952	deb01a94.daz	184	deb03c94.enz	328	deb07a94.esz
1048	deb01a94.dez	192	deb03c94.esz	344	deb07a94.frz
1176	deb01a94.elz	200	deb03c94.frz	328	deb07a94.itz
920	deb01a94.enz	192	deb03c94.itz	328	deb07a94.nlz
1008	deb01a94.esz	192	deb03c94.nez	336	deb07a94.ptz
1032	deb01a94.frz	192	deb03c94.poz	944	deb09a93.daz
960	deb01a94.itz	984	deb04a93.daz	920	deb09a93.enz
1000	deb01a94.nez	960	deb04a93.enz	1000	deb09a93.esz
1000	deb01a94.poz	1056	deb04a93.esz	1000	deb09a93.nez
992	deb02a94.daz	1048	deb04a93.nez	1008	deb09a93.poz
1104	deb02a94.dez	1048	deb04a93.poz	192	deb09b93.daz
1216	deb02a94.elz	976	deb04a94.daz	192	deb09b93.enz
960	deb02a94.enz	1072	deb04a94.dez	200	deb09b93.esz
1048	deb02a94.esz	1192	deb04a94.elz	208	deb09b93.frz
1080	deb02a94.frz	944	deb04a94.enz	200	deb09b93.nez
1016	deb02a94.itz	1024	deb04a94.esz	208	deb09b93.poz
1048	deb02a94.nez	1072	deb04a94.frz	200	deb10a93.daz
1056	deb02a94.poz	1008	deb04a94.itz	200	deb10a93.enz
200	deb02b94.daz	1016	deb04a94.nez	208	deb10a93.esz
208	deb02b94.dez	1040	deb04a94.poz	224	deb10a93.frz
240	deb02b94.elz	992	deb05a93.daz	216	deb10a93.nez
192	deb02b94.enz	968	deb05a93.enz	216	deb10a93.poz
208	deb02b94.esz	1056	deb05a93.esz	992	deb10b93.daz
208	deb02b94.frz	1048	deb05a93.nez	968	deb10b93.enz
200	deb02b94.itz	1064	deb05a93.poz	1056	deb10b93.esz
208	deb02b94.nez	960	deb05a94.daz	1088	deb10b93.frz
200	deb02b94.poz	1064	deb05a94.dez	1040	deb10b93.nez
152	deb03a94.daz	1168	deb05a94.elz	1064	deb10b93.poz
168	deb03a94.dez	928	deb05a94.enz	936	deb11a93.daz
184	deb03a94.elz	1008	deb05a94.esz	912	deb11a93.enz
144	deb03a94.enz	1040	deb05a94.frz	984	deb11a93.esz
160	deb03a94.esz	992	deb05a94.itz	1008	deb11a93.frz
168	deb03a94.frz	1008	deb05a94.nez	992	deb11a93.nez
160	deb03a94.itz	1016	deb05a94.poz	984	deb11a93.poz
160	deb03a94.nez	1008	deb06a93.daz	208	deb12a93.daz
160	deb03a94.poz	992	deb06a93.enz	208	deb12a93.enz
760	deb03b94.daz	1080	deb06a93.esz	208	deb12a93.esz
840	deb03b94.dez	1072	deb06a93.nez	216	deb12a93.frz
936	deb03b94.elz	1080	deb06a93.poz	216	deb12a93.nez
744	deb03b94.enz	1072	deb07a93.daz	216	deb12a93.poz
808	deb03b94.esz	1048	deb07a93.enz	960	deb12b93.daz
840	deb03b94.frz	1144	deb07a93.esz	928	deb12b93.enz
808	deb03b94.itz	1144	deb07a93.nez	1008	deb12b93.esz
808	deb03b94.nez	1144	deb07a93.poz	1024	deb12b93.frz
816	deb03b94.poz	320	deb07a94.daz	1008	deb12b93.nez
184	deb03c94.daz	344	deb07a94.dez	1008	deb12b93.poz
208	deb03c94.dez	392	deb07a94.elz		

MLCC-RELEASE/MLCC-DEBATES/orig/diskettes: total 6

1	dadisks	3	dedisks	2	eldisks

MLCC-RELEASE/MLCC-DEBATES/orig/diskettes/dadisks: total 5037

392	ep3_413.da.Z	496	ep3_417.da1.Z	176	ep3_419.da2.Z
480	ep3_414.da1.Z	144	ep3_417.da2.Z	504	ep3_420.da1.Z
168	ep3_414.da2.Z	488	ep3_418.da1.Z	216	ep3_420.da2.Z
85	ep3_415.da.Z	184	ep3_418.da2.Z	256	ep3_421.da1.Z
512	ep3_416.da.Z	496	ep3_419.da1.Z	440	ep3_421.da2.Z

MLCC-RELEASE/MLCC-DEBATES/orig/diskettes/dedisks: total 18176

11	00README	41	3-421.500.Z	320	3-431.400.Z
3	3-413.100.Z	2	3-421.600.Z	39	3-431.500.Z
17	3-413.200.Z	136	3-423.100.Z	2	3-431.600.Z
224	3-413.300.Z	208	3-423.200.Z	128	3-432.100.Z
136	3-413.400.Z	328	3-423.300.Z	160	3-432.200.Z
63	3-413.500.Z	112	3-423.400.Z	280	3-432.300.Z
2	3-413.600.Z	71	3-423.500.Z	160	3-432.400.Z
81	3-414.100.Z	2	3-423.600.Z	78	3-432.500.Z
184	3-414.200.Z	136	3-424.100.Z	2	3-432.600.Z
280	3-414.300.Z	184	3-424.200.Z	128	3-433.100.Z
152	3-414.400.Z	280	3-424.300.Z	200	3-433.200.Z
35	3-414.500.Z	168	3-424.400.Z	296	3-433.300.Z
2	3-414.600.Z	69	3-424.500.Z	176	3-433.400.Z
94	3-415.100.Z	3	3-424.600.Z	69	3-433.500.Z
1	3-415.200.Z	128	3-425.100.Z	2	3-433.600.Z
168	3-416.200.Z	184	3-425.200.Z	78	3-434.100.Z
248	3-416.300.Z	296	3-425.300.Z	184	3-434.200.Z
160	3-416.400.Z	168	3-425.400.Z	272	3-434.300.Z
35	3-416.500.Z	73	3-425.500.Z	168	3-434.400.Z
2	3-416.600.Z	2	3-425.600.Z	51	3-434.500.Z
120	3-417.100.Z	128	3-426.100.Z	2	3-434.600.Z
152	3-417.200.Z	192	3-426.200.Z	144	3-435.100.Z
280	3-417.300.Z	288	3-426.300.Z	1	3-435.200.Z
144	3-417.400.Z	176	3-426.400.Z	79	3-436.100.Z
30	3-417.500.Z	73	3-426.500.Z	74	3-436.200.Z
2	3-417.600.Z	2	3-426.600.Z	1	3-436.300.Z
128	3-418.100.Z	120	3-427.100.Z	120	3-437.100.Z
176	3-418.200.Z	192	3-427.200.Z	168	3-437.200.Z
272	3-418.300.Z	496	3-427.300.Z	304	3-437.300.Z
144	3-418.400.Z	168	3-427.400.Z	152	3-437.400.Z
67	3-418.500.Z	50	3-427.500.Z	53	3-437.500.Z
3	3-418.600.Z	2	3-427.600.Z	2	3-437.600.Z
120	3-419.100.Z	128	3-428.100.Z	120	3-438.100.Z
168	3-419.200.Z	1	3-428.600.Z	168	3-438.200.Z
296	3-419.300.Z	192	3-429.200.Z	232	3-438.300.Z
152	3-419.400.Z	328	3-429.300.Z	168	3-438.400.Z
55	3-419.500.Z	168	3-429.400.Z	66	3-438.500.Z
2	3-419.600.Z	76	3-429.500.Z	2	3-438.600.Z
120	3-420.100.Z	2	3-429.600.Z	94	3-439.100.Z
184	3-420.200.Z	120	3-430.100.Z	62	3-439.200.Z
288	3-420.300.Z	192	3-430.200.Z	1	3-439.300.Z
168	3-420.400.Z	320	3-430.300.Z	112	3-440.100.Z
76	3-420.500.Z	128	3-430.400.Z	184	3-440.200.Z
3	3-420.600.Z	38	3-430.500.Z	240	3-440.300.Z
120	3-421.100.Z	2	3-430.600.Z	168	3-440.400.Z
176	3-421.200.Z	128	3-431.100.Z	75	3-440.500.Z
304	3-421.300.Z	176	3-431.200.Z	2	3-440.600.Z
168	3-421.400.Z	296	3-431.300.Z		

MLCC-RELEASE/MLCC-DEBATES/orig/diskettes/eldisks: total 18047

7	00README	352	ep3_423.gr1.Z	504	ep3_432.gr1.Z
464	ep3_413.gr.Z	512	ep3_423.gr2.Z	320	ep3_432.gr2.Z
472	ep3_414.gr1.Z	536	ep3_424.gr1.Z	408	ep3_433.gr1.Z
280	ep3_414.gr2.Z	312	ep3_424.gr2.Z	464	ep3_433.gr2.Z
112	ep3_415.gr.Z	544	ep3_425.gr1.Z	544	ep3_434.gr1.Z
352	ep3_416.gr1.Z	320	ep3_425.gr2.Z	232	ep3_434.gr2.Z
288	ep3_416.gr2.Z	528	ep3_426.gr1.Z	152	ep3_435.gr.Z
504	ep3_417.gr1.Z	336	ep3_426.gr2.Z	160	ep3_436.gr.Z
248	ep3_417.gr2.Z	536	ep3_427.gr1.Z	504	ep3_437.gr1.Z
496	ep3_418.gr1.Z	264	ep3_427.gr2.Z	304	ep3_437.gr2.Z
296	ep3_418.gr2.Z	136	ep3_428.gr.Z	480	ep3_438.gr1.Z
520	ep3_419.gr1.Z	440	ep3_429.gr1.Z	272	ep3_438.gr2.Z
280	ep3_419.gr2.Z	344	ep3_429.gr2.Z	168	ep3_439.gr.Z
536	ep3_420.gr1.Z	536	ep3_430.gr1.Z	512	ep3_440.gr1.Z
312	ep3_420.gr2.Z	280	ep3_430.gr2.Z	256	ep3_440.gr2.Z
520	ep3_421.gr1.Z	520	ep3_431.gr1.Z		
288	ep3_421.gr2.Z	296	ep3_431.gr2.Z		

MLCC-RELEASE/MLCC-DEBATES/orig/newtapes: total 9033

1	00README	2	n1.Z	2	n28.Z	1	n46.Z	2	p17.Z
2	f1.Z	2	n10.Z	2	n29.Z	120	n47.Z	2	p18.Z
2	f10.Z	1	n11.Z	1	n3.Z	160	n48.Z	1	p19.Z
1	f11.Z	81	n12.Z	1	n30.Z	296	n49.Z	2	p2.Z
128	f12.Z	184	n13.Z	128	n31.Z	17	n5.Z	128	p20.Z
176	f13.Z	272	n14.Z	168	n32.Z	144	n50.Z	192	p21.Z
288	f14.Z	152	n15.Z	280	n33.Z	53	n51.Z	328	p22.Z
160	f15.Z	36	n16.Z	136	n34.Z	432	n54.Z	112	p23.Z
70	f16.Z	1	n17.Z	29	n35.Z	240	n6.Z	71	p24.Z
2	f2.Z	1	n18.Z	2	n36.Z	136	n7.Z	85	p27.Z
85	f20.Z	1	n19.Z	2	n37.Z	62	n8.Z	1	p28.Z
28	f22.Z	2	n2.Z	1	n38.Z	2	n9.Z	1	p3.Z
70	f23.Z	92	n20.Z	128	n39.Z	2	p1.Z	1	p30.Z
1	f3.Z	2	n21.Z	3	n4.Z	2	p10.Z	1	p31.Z
136	f4.Z	2	n22.Z	168	n40.Z	1	p11.Z	120	p4.Z
184	f5.Z	1	n23.Z	272	n41.Z	120	p12.Z	176	p5.Z
280	f6.Z	168	n24.Z	144	n42.Z	168	p13.Z	288	p6.Z
160	f7.Z	248	n25.Z	66	n43.Z	304	p14.Z	168	p7.Z
70	f8.Z	152	n26.Z	2	n44.Z	160	p15.Z	76	p8.Z
2	f9.Z	35	n27.Z	2	n45.Z	38	p16.Z	2	p9.Z

MLCC-RELEASE/MLCC-DEBATES/orig/tapesascii: total 54136

38632	tape1.Z	4368	tape3.Z	2216	tape5.Z
4064	tape2.Z	3808	tape4.Z	1048	tape6.Z

MLCC-RELEASE/MLCC-DEBATES/orig/tapesmopas: total 81973

1	00README	1472	380.5.Z	1592	385.2.Z	1496	388.2.Z	2112	391.7.Z
888	378.1.Z	1456	381.1.Z	1608	385.3.Z	248	388.3.Z	1	391.8.Z
1392	378.2.Z	1568	381.2.Z	1496	385.4.Z	1480	388.4.Z	1488	392.1.Z
1352	378.3.Z	1512	381.3.Z	248	385.5.Z	1416	389.1.Z	912	392.10.Z
1416	378.4.Z	1608	381.4.Z	1456	386.1.Z	224	389.2.Z	1024	392.11.Z
1464	378.5.Z	1456	382.1.Z	1464	386.2.Z	1584	389.3.Z	1	392.12.Z
880	379.1.Z	1560	382.2.Z	1480	386.3.Z	12	389.6.Z	1792	392.3.Z
1400	379.2.Z	1536	382.3.Z	1384	386.4.Z	1344	390.1.Z	264	392.4.Z
1360	379.3.Z	1608	382.4.Z	216	386.5.Z	9	390.4.Z	18	392.5.Z
1416	379.4.Z	2600	383.1.Z	1576	387.1.Z	1072	390.6.Z	75	392.7.Z
1472	379.5.Z	1504	384.1.Z	1592	387.2.Z	1488	390.7.Z	86	392.8.Z
1344	380.1.Z	1472	384.2.Z	1608	387.3.Z	1472	391.1.Z		
1344	380.2.Z	1480	384.3.Z	1504	387.4.Z	17	391.14.Z		
1448	380.3.Z	1512	384.4.Z	248	387.5.Z	752	391.3.Z		
1392	380.4.Z	1576	385.1.Z	1624	388.1.Z	1	391.4.Z		

MLCC-RELEASE/MLCC-DEBATES/prep: total 9

2	add-headers	2	new-tei-header	1	tapesmopas
1	bernou	1	new-tei-trailer		
1	diskettes	1	tapesascii		

MLCC-RELEASE/MLCC-DEBATES/prep/bernou: total 226

1	00README	1	final	1	unzipall.bat
1	checksgml.bat	1	removebin.bat	1	unzipcheck.bat
93	eurobern.c	6	rename.bat		
120	eurobern.exe	1	sgmltag.bat		

MLCC-RELEASE/MLCC-DEBATES/prep/diskettes: total 1

1	copy.bat.Z

MLCC-RELEASE/MLCC-DEBATES/prep/tapesascii: total 80

1	00README	16	dos2unix.exe	1	final
1	asciitosgml.bat	19	euroasc.c	1	rentape1.bat
1	chksgml.bat	40	euroasc.exe		

MLCC-RELEASE/MLCC-DEBATES/prep/tapesmopas: total 186

1	00README	1	conv3.bat	82	europarl.c
1	conv1.bat	2	conv4.bat	96	europarl.exe
1	conv2.bat	1	copyfiles.bat	1	final

C.3.2 Dutch Newspaper Corpus

MLCC-RELEASE/MLCC-DUTCH: total 28

1	00COPYRIGHT	7	doc.body	1	licence
2	00README	1	doc.tex	1	orig
1	data	13	editdecl.txt	1	prep

MLCC-RELEASE/MLCC-DUTCH/data: total 21269

352	DUTCH.HFD.sgm.gz	720	ed9209.sgm.gz	816	ed9306.sgm.gz
29	ENGLISH.HFD.sgm.gz	832	ed9210.sgm.gz	704	ed9307.sgm.gz
784	ed9201.sgm.gz	800	ed9211.sgm.gz	744	ed9308.sgm.gz
768	ed9202.sgm.gz	792	ed9212.sgm.gz	768	ed9309.sgm.gz
856	ed9203.sgm.gz	856	ed9213.sgm.gz	840	ed9310.sgm.gz
880	ed9204.sgm.gz	688	ed9301.sgm.gz	824	ed9311.sgm.gz
824	ed9205.sgm.gz	808	ed9302.sgm.gz	976	ed9312.sgm.gz
872	ed9206.sgm.gz	784	ed9303.sgm.gz	808	ed9313.sgm.gz
816	ed9207.sgm.gz	808	ed9304.sgm.gz		
712	ed9208.sgm.gz	808	ed9305.sgm.gz		

MLCC-RELEASE/MLCC-DUTCH/licence: total 11

5	agr-ed-multext-hfd.tex	6	non-disclosure-agr-hfd.tex

MLCC-RELEASE/MLCC-DUTCH/orig: total 21271

2	00README	712	Ed9208.Lst.gz	808	Ed9305.Lst.gz
352	DUTCH.HFD.gz	728	Ed9209.Lst.gz	816	Ed9306.Lst.gz
29	ENGLISH.HFD.gz	832	Ed9210.Lst.gz	704	Ed9307.Lst.gz
784	Ed9201.Lst.gz	800	Ed9211.Lst.gz	744	Ed9308.Lst.gz
768	Ed9202.Lst.gz	792	Ed9212.Lst.gz	768	Ed9309.Lst.gz
856	Ed9203.Lst.gz	856	Ed9213.Lst.gz	840	Ed9310.Lst.gz
880	Ed9204.Lst.gz	688	Ed9301.Lst.gz	816	Ed9311.Lst.gz
824	Ed9205.Lst.gz	808	Ed9302.Lst.gz	976	Ed9312.Lst.gz
872	Ed9206.Lst.gz	784	Ed9303.Lst.gz	808	Ed9313.Lst.gz
816	Ed9207.Lst.gz	808	Ed9304.Lst.gz		

MLCC-RELEASE/MLCC-DUTCH/prep: total 32

1	add-headers	1	prep2.pl	1	prepE.pl
4	dutch_dtd	1	prepA.pl	4	script
3	new-tei-header	4	prepB.pl	1	script2
1	new-tei-trailer	6	prepC.pl	3	script3
1	prep.pl	1	prepD.pl		

C.3.3 English Newspaper Corpus

MLCC-RELEASE/MLCC-ENGLISH: total 22

1	00COPYRIGHT	11	doc.body	1	licence
2	00README	1	doc.tex	2	orig
2	data	1	editdecl.txt	1	prep

MLCC-RELEASE/MLCC-ENGLISH/data: total 68319

888	tape1aa.sgm.gz	544	tape1bj.sgm.gz	888	tape2ar.sgm.gz
880	tape1ab.sgm.gz	880	tape1bk.sgm.gz	824	tape2ba.sgm.gz
824	tape1ac.sgm.gz	864	tape1bl.sgm.gz	904	tape2bb.sgm.gz
896	tape1ad.sgm.gz	872	tape1bm.sgm.gz	848	tape2bc.sgm.gz
888	tape1ae.sgm.gz	872	tape1bn.sgm.gz	872	tape2bd.sgm.gz
840	tape1af.sgm.gz	840	tape1bo.sgm.gz	888	tape2be.sgm.gz
912	tape1ag.sgm.gz	856	tape1bp.sgm.gz	864	tape2bf.sgm.gz
824	tape1ah.sgm.gz	880	tape1bq.sgm.gz	832	tape2bg.sgm.gz
872	tape1ai.sgm.gz	872	tape1br.sgm.gz	872	tape2bh.sgm.gz
896	tape1aj.sgm.gz	664	tape1bs.sgm.gz	720	tape2bi.sgm.gz
920	tape1ak.sgm.gz	896	tape2aa.sgm.gz	864	tape2bj.sgm.gz
848	tape1al.sgm.gz	824	tape2ab.sgm.gz	904	tape2bk.sgm.gz
888	tape1am.sgm.gz	864	tape2ac.sgm.gz	824	tape2bl.sgm.gz
95	tape1an.sgm.gz	848	tape2ad.sgm.gz	888	tape2bm.sgm.gz
888	tape1ao.sgm.gz	848	tape2ae.sgm.gz	856	tape2bn.sgm.gz
848	tape1ap.sgm.gz	888	tape2af.sgm.gz	848	tape2bo.sgm.gz
888	tape1aq.sgm.gz	912	tape2ag.sgm.gz	920	tape2bp.sgm.gz
864	tape1ar.sgm.gz	856	tape2ah.sgm.gz	880	tape2bq.sgm.gz
896	tape1ba.sgm.gz	888	tape2ai.sgm.gz	776	tape2br.sgm.gz
896	tape1bb.sgm.gz	856	tape2aj.sgm.gz	880	tape3aa.sgm.gz
840	tape1bc.sgm.gz	896	tape2ak.sgm.gz	880	tape3ab.sgm.gz
904	tape1bd.sgm.gz	904	tape2al.sgm.gz	832	tape3ac.sgm.gz
912	tape1be.sgm.gz	904	tape2am.sgm.gz	856	tape3ad.sgm.gz
872	tape1bf.sgm.gz	624	tape2an.sgm.gz	888	tape3ae.sgm.gz
896	tape1bg.sgm.gz	888	tape2ao.sgm.gz	888	tape3af.sgm.gz
832	tape1bh.sgm.gz	832	tape2ap.sgm.gz	872	tape3ag.sgm.gz
904	tape1bi.sgm.gz	896	tape2aq.sgm.gz	272	tape3ah.sgm.gz

MLCC-RELEASE/MLCC-ENGLISH/licence: total 11

6	edin-ft-agreement.tex	5	edin-user-agreement.tex

MLCC-RELEASE/MLCC-ENGLISH/orig: total 67013

872	tape1aa.gz	848	tape1ar.gz	864	tape1bq.gz	872	tape2ao.gz	840	tape2bn.gz
864	tape1ab.gz	880	tape1ba.gz	856	tape1br.gz	816	tape2ap.gz	832	tape2bo.gz
808	tape1ac.gz	872	tape1bb.gz	656	tape1bs.gz	880	tape2aq.gz	904	tape2bp.gz
872	tape1ad.gz	824	tape1bc.gz	880	tape2aa.gz	872	tape2ar.gz	864	tape2bq.gz
872	tape1ae.gz	888	tape1bd.gz	800	tape2ab.gz	808	tape2ba.gz	760	tape2br.gz
824	tape1af.gz	896	tape1be.gz	848	tape2ac.gz	888	tape2bb.gz	864	tape3aa.gz
896	tape1ag.gz	856	tape1bf.gz	832	tape2ad.gz	832	tape2bc.gz	864	tape3ab.gz
808	tape1ah.gz	880	tape1bg.gz	832	tape2ae.gz	856	tape2bd.gz	816	tape3ac.gz
856	tape1ai.gz	816	tape1bh.gz	872	tape2af.gz	872	tape2be.gz	840	tape3ad.gz
880	tape1aj.gz	888	tape1bi.gz	896	tape2ag.gz	848	tape2bf.gz	872	tape3ae.gz
904	tape1ak.gz	536	tape1bj.gz	840	tape2ah.gz	816	tape2bg.gz	872	tape3af.gz
832	tape1al.gz	856	tape1bk.gz	872	tape2ai.gz	856	tape2bh.gz	856	tape3ag.gz
864	tape1am.gz	848	tape1bl.gz	840	tape2aj.gz	704	tape2bi.gz	264	tape3ah.gz
93	tape1an.gz	848	tape1bm.gz	880	tape2ak.gz	848	tape2bj.gz		
872	tape1ao.gz	856	tape1bn.gz	888	tape2al.gz	888	tape2bk.gz		
832	tape1ap.gz	816	tape1bo.gz	888	tape2am.gz	808	tape2bl.gz		
872	tape1aq.gz	840	tape1bp.gz	616	tape2an.gz	872	tape2bm.gz		

MLCC-RELEASE/MLCC-ENGLISH/prep: total 35

2	FT-TAPE-todo	1	convert4.perl	4	script
3	convert.perl.old	1	del_blank.sed	7	script2
3	convert1.perl	4	histogram1	1	transform1.perl
1	convert2.perl	3	new-tei-header	1	transform2.perl
3	convert3.perl	1	new-tei-trailer		

C.3.4 French Newspaper Corpus

MLCC-RELEASE/MLCC-FRENCH: total 30

2	00COPYRIGHT	9	doc.body	1	licence
2	00README	1	doc.tex	2	orig
4	data	8	editdecl.txt	1	prep

MLCC-RELEASE/MLCC-FRENCH/data: total 20996

328	fr01A1292.01.sgm.gz	336	fr01A1293.02.sgm.gz
320	fr01A1292.02.sgm.gz	336	fr01A1293.03.sgm.gz
336	fr01A1292.03.sgm.gz	328	fr01A1293.04.sgm.gz
344	fr01A1292.04.sgm.gz	344	fr01A1293.05.sgm.gz
336	fr01A1292.05.sgm.gz	336	fr01A1293.06.sgm.gz
336	fr01A1292.06.sgm.gz	344	fr01A1293.07.sgm.gz
336	fr01A1292.07.sgm.gz	352	fr01A1293.08.sgm.gz
336	fr01A1292.08.sgm.gz	336	fr01A1293.09.sgm.gz
344	fr01A1292.09.sgm.gz	344	fr01A1293.10.sgm.gz
328	fr01A1292.10.sgm.gz	336	fr01A1293.11.sgm.gz
336	fr01A1292.11.sgm.gz	344	fr01A1293.12.sgm.gz
336	fr01A1292.12.sgm.gz	328	fr01A1293.13.sgm.gz
344	fr01A1292.13.sgm.gz	336	fr01A1293.14.sgm.gz
344	fr01A1292.14.sgm.gz	344	fr01A1293.15.sgm.gz
352	fr01A1292.15.sgm.gz	336	fr01A1293.16.sgm.gz
328	fr01A1292.16.sgm.gz	336	fr01A1293.17.sgm.gz
320	fr01A1292.17.sgm.gz	344	fr01A1293.18.sgm.gz
320	fr01A1292.18.sgm.gz	328	fr01A1293.19.sgm.gz
352	fr01A1292.19.sgm.gz	328	fr01A1293.20.sgm.gz
336	fr01A1292.20.sgm.gz	352	fr01A1293.21.sgm.gz
336	fr01A1292.21.sgm.gz	368	fr01A1293.22.sgm.gz
328	fr01A1292.22.sgm.gz	384	fr01A1293.23.sgm.gz
336	fr01A1292.23.sgm.gz	400	fr01A1293.24.sgm.gz
336	fr01A1292.24.sgm.gz	392	fr01A1293.25.sgm.gz
336	fr01A1292.25.sgm.gz	376	fr01A1293.26.sgm.gz
328	fr01A1292.26.sgm.gz	384	fr01A1293.27.sgm.gz
336	fr01A1292.27.sgm.gz	384	fr01A1293.28.sgm.gz
328	fr01A1292.28.sgm.gz	384	fr01A1293.29.sgm.gz
320	fr01A1292.29.sgm.gz	384	fr01A1293.30.sgm.gz
304	fr01A1292.30.sgm.gz	376	fr01A1293.31.sgm.gz
336	fr01A1293.01.sgm.gz	60	fr01A1293.32.sgm.gz

MLCC-RELEASE/MLCC-FRENCH/licence: total 10

5	edin-user-agreement.tex	5	ldc-user-agreement.tex

MLCC-RELEASE/MLCC-FRENCH/orig: total 21003

328	fr01A1292.01.gz	328	fr01A1292.22.gz	328	fr01A1293.13.gz
320	fr01A1292.02.gz	336	fr01A1292.23.gz	336	fr01A1293.14.gz
336	fr01A1292.03.gz	336	fr01A1292.24.gz	344	fr01A1293.15.gz
344	fr01A1292.04.gz	336	fr01A1292.25.gz	336	fr01A1293.16.gz
336	fr01A1292.05.gz	328	fr01A1292.26.gz	336	fr01A1293.17.gz
336	fr01A1292.06.gz	336	fr01A1292.27.gz	344	fr01A1293.18.gz
336	fr01A1292.07.gz	328	fr01A1292.28.gz	328	fr01A1293.19.gz
336	fr01A1292.08.gz	320	fr01A1292.29.gz	336	fr01A1293.20.gz
344	fr01A1292.09.gz	304	fr01A1292.30.gz	352	fr01A1293.21.gz
328	fr01A1292.10.gz	336	fr01A1293.01.gz	368	fr01A1293.22.gz
336	fr01A1292.11.gz	336	fr01A1293.02.gz	384	fr01A1293.23.gz
336	fr01A1292.12.gz	336	fr01A1293.03.gz	400	fr01A1293.24.gz
344	fr01A1292.13.gz	328	fr01A1293.04.gz	392	fr01A1293.25.gz
344	fr01A1292.14.gz	344	fr01A1293.05.gz	376	fr01A1293.26.gz
352	fr01A1292.15.gz	336	fr01A1293.06.gz	384	fr01A1293.27.gz
328	fr01A1292.16.gz	344	fr01A1293.07.gz	384	fr01A1293.28.gz
320	fr01A1292.17.gz	352	fr01A1293.08.gz	384	fr01A1293.29.gz
320	fr01A1292.18.gz	336	fr01A1293.09.gz	384	fr01A1293.30.gz
352	fr01A1292.19.gz	344	fr01A1293.10.gz	376	fr01A1293.31.gz
336	fr01A1292.20.gz	336	fr01A1293.11.gz	59	fr01A1293.32.gz
336	fr01A1292.21.gz	344	fr01A1293.12.gz		

MLCC-RELEASE/MLCC-FRENCH/prep: total 28

1	add-headers	2	prep.perl	3	script
1	final	1	prep2.perl	3	script2
1	fixids.perl	3	prep3.perl	3	structure
2	lemonde.dtd	1	prep4.perl	1	xxx.perl
3	new-tei-header	1	run		
1	new-tei-trailer	1	run2		

C.3.5 German Newspaper Corpus

MLCC-RELEASE/MLCC-GERMAN: total 32

1	00COPYRIGHT	6	doc.body	1	licence
5	00README	1	doc.tex	5	orig
11	data	1	editdecl.txt	1	prep

MLCC-RELEASE/MLCC-GERMAN/data: total 94472

456	hb001.sgm.gz	464	hb071.sgm.gz	464	hb141.sgm.gz
440	hb002.sgm.gz	464	hb072.sgm.gz	464	hb142.sgm.gz
456	hb003.sgm.gz	456	hb073.sgm.gz	456	hb143.sgm.gz
456	hb004.sgm.gz	464	hb074.sgm.gz	448	hb144.sgm.gz
456	hb005.sgm.gz	456	hb075.sgm.gz	448	hb145.sgm.gz
456	hb006.sgm.gz	464	hb076.sgm.gz	448	hb146.sgm.gz
456	hb007.sgm.gz	464	hb077.sgm.gz	456	hb147.sgm.gz
456	hb008.sgm.gz	456	hb078.sgm.gz	448	hb148.sgm.gz
456	hb009.sgm.gz	464	hb079.sgm.gz	448	hb149.sgm.gz
448	hb010.sgm.gz	456	hb080.sgm.gz	440	hb150.sgm.gz
456	hb011.sgm.gz	464	hb081.sgm.gz	448	hb151.sgm.gz
456	hb012.sgm.gz	464	hb082.sgm.gz	448	hb152.sgm.gz
448	hb013.sgm.gz	464	hb083.sgm.gz	456	hb153.sgm.gz
456	hb014.sgm.gz	464	hb084.sgm.gz	448	hb154.sgm.gz
456	hb015.sgm.gz	472	hb085.sgm.gz	464	hb155.sgm.gz
456	hb016.sgm.gz	464	hb086.sgm.gz	456	hb156.sgm.gz
464	hb017.sgm.gz	464	hb087.sgm.gz	456	hb157.sgm.gz
456	hb018.sgm.gz	464	hb088.sgm.gz	448	hb158.sgm.gz
448	hb019.sgm.gz	456	hb089.sgm.gz	400	hb159.sgm.gz
448	hb020.sgm.gz	456	hb090.sgm.gz	392	hb160.sgm.gz
448	hb021.sgm.gz	464	hb091.sgm.gz	392	hb161.sgm.gz
456	hb022.sgm.gz	464	hb092.sgm.gz	392	hb162.sgm.gz
464	hb023.sgm.gz	464	hb093.sgm.gz	392	hb163.sgm.gz
456	hb024.sgm.gz	456	hb094.sgm.gz	392	hb164.sgm.gz
456	hb025.sgm.gz	456	hb095.sgm.gz	392	hb165.sgm.gz
456	hb026.sgm.gz	464	hb096.sgm.gz	392	hb166.sgm.gz
456	hb027.sgm.gz	456	hb097.sgm.gz	384	hb167.sgm.gz
456	hb028.sgm.gz	464	hb098.sgm.gz	392	hb168.sgm.gz
456	hb029.sgm.gz	464	hb099.sgm.gz	400	hb169.sgm.gz
456	hb030.sgm.gz	464	hb100.sgm.gz	392	hb170.sgm.gz
456	hb031.sgm.gz	456	hb101.sgm.gz	392	hb171.sgm.gz
456	hb032.sgm.gz	456	hb102.sgm.gz	392	hb172.sgm.gz
456	hb033.sgm.gz	456	hb103.sgm.gz	392	hb173.sgm.gz
456	hb034.sgm.gz	456	hb104.sgm.gz	400	hb174.sgm.gz
456	hb035.sgm.gz	464	hb105.sgm.gz	392	hb175.sgm.gz
456	hb036.sgm.gz	448	hb106.sgm.gz	456	hb176.sgm.gz
456	hb037.sgm.gz	472	hb107.sgm.gz	456	hb177.sgm.gz
456	hb038.sgm.gz	456	hb108.sgm.gz	464	hb178.sgm.gz
456	hb039.sgm.gz	456	hb109.sgm.gz	456	hb179.sgm.gz
440	hb040.sgm.gz	456	hb110.sgm.gz	464	hb180.sgm.gz
464	hb041.sgm.gz	456	hb111.sgm.gz	464	hb181.sgm.gz
456	hb042.sgm.gz	456	hb112.sgm.gz	448	hb182.sgm.gz
464	hb043.sgm.gz	480	hb113.sgm.gz	456	hb183.sgm.gz
464	hb044.sgm.gz	400	hb114.sgm.gz	456	hb184.sgm.gz
464	hb045.sgm.gz	400	hb115.sgm.gz	464	hb185.sgm.gz
464	hb046.sgm.gz	400	hb116.sgm.gz	456	hb186.sgm.gz
472	hb047.sgm.gz	400	hb117.sgm.gz	448	hb187.sgm.gz
456	hb048.sgm.gz	400	hb118.sgm.gz	448	hb188.sgm.gz
472	hb049.sgm.gz	400	hb119.sgm.gz	456	hb189.sgm.gz
464	hb050.sgm.gz	408	hb120.sgm.gz	456	hb190.sgm.gz
464	hb051.sgm.gz	400	hb121.sgm.gz	456	hb191.sgm.gz
464	hb052.sgm.gz	400	hb122.sgm.gz	456	hb192.sgm.gz
472	hb053.sgm.gz	392	hb123.sgm.gz	480	hb193.sgm.gz
456	hb054.sgm.gz	408	hb124.sgm.gz	464	hb194.sgm.gz
464	hb055.sgm.gz	400	hb125.sgm.gz	472	hb195.sgm.gz
456	hb056.sgm.gz	400	hb126.sgm.gz	456	hb196.sgm.gz
464	hb057.sgm.gz	448	hb127.sgm.gz	472	hb197.sgm.gz
464	hb058.sgm.gz	464	hb128.sgm.gz	464	hb198.sgm.gz
448	hb059.sgm.gz	456	hb129.sgm.gz	472	hb199.sgm.gz
488	hb060.sgm.gz	456	hb130.sgm.gz	456	hb200.sgm.gz
464	hb061.sgm.gz	464	hb131.sgm.gz	472	hb201.sgm.gz
448	hb062.sgm.gz	464	hb132.sgm.gz	456	hb202.sgm.gz
456	hb063.sgm.gz	464	hb133.sgm.gz	464	hb203.sgm.gz
448	hb064.sgm.gz	472	hb134.sgm.gz	472	hb204.sgm.gz
456	hb065.sgm.gz	464	hb135.sgm.gz	464	hb205.sgm.gz
456	hb066.sgm.gz	464	hb136.sgm.gz	464	hb206.sgm.gz
456	hb067.sgm.gz	464	hb137.sgm.gz	464	hb207.sgm.gz
456	hb068.sgm.gz	464	hb138.sgm.gz	464	hb208.sgm.gz
456	hb069.sgm.gz	464	hb139.sgm.gz	464	hb209.sgm.gz
456	hb070.sgm.gz	456	hb140.sgm.gz	456	hb210.sgm.gz

MLCC-RELEASE/MLCC-GERMAN/licence: total 1

1	00README

MLCC-RELEASE/MLCC-GERMAN/orig: total 106512

512	hb.001.gz	520	hb.043.gz	528	hb.085.gz	504	hb.127.gz	448	hb.169.gz
496	hb.002.gz	520	hb.044.gz	528	hb.086.gz	520	hb.128.gz	440	hb.170.gz
512	hb.003.gz	528	hb.045.gz	520	hb.087.gz	520	hb.129.gz	440	hb.171.gz
520	hb.004.gz	520	hb.046.gz	520	hb.088.gz	520	hb.130.gz	440	hb.172.gz
520	hb.005.gz	528	hb.047.gz	512	hb.089.gz	528	hb.131.gz	440	hb.173.gz
512	hb.006.gz	520	hb.048.gz	520	hb.090.gz	528	hb.132.gz	448	hb.174.gz
520	hb.007.gz	528	hb.049.gz	520	hb.091.gz	528	hb.133.gz	440	hb.175.gz
512	hb.008.gz	528	hb.050.gz	520	hb.092.gz	528	hb.134.gz	512	hb.176.gz
520	hb.009.gz	520	hb.051.gz	528	hb.093.gz	520	hb.135.gz	512	hb.177.gz
512	hb.010.gz	520	hb.052.gz	520	hb.094.gz	528	hb.136.gz	520	hb.178.gz
512	hb.011.gz	528	hb.053.gz	512	hb.095.gz	520	hb.137.gz	520	hb.179.gz
512	hb.012.gz	520	hb.054.gz	528	hb.096.gz	520	hb.138.gz	520	hb.180.gz
512	hb.013.gz	520	hb.055.gz	520	hb.097.gz	520	hb.139.gz	520	hb.181.gz
512	hb.014.gz	512	hb.056.gz	528	hb.098.gz	520	hb.140.gz	512	hb.182.gz
512	hb.015.gz	520	hb.057.gz	520	hb.099.gz	528	hb.141.gz	520	hb.183.gz
512	hb.016.gz	520	hb.058.gz	528	hb.100.gz	520	hb.142.gz	512	hb.184.gz
520	hb.017.gz	504	hb.059.gz	512	hb.101.gz	512	hb.143.gz	520	hb.185.gz
520	hb.018.gz	536	hb.060.gz	512	hb.102.gz	512	hb.144.gz	512	hb.186.gz
512	hb.019.gz	520	hb.061.gz	520	hb.103.gz	504	hb.145.gz	512	hb.187.gz
512	hb.020.gz	504	hb.062.gz	520	hb.104.gz	504	hb.146.gz	504	hb.188.gz
504	hb.021.gz	512	hb.063.gz	520	hb.105.gz	512	hb.147.gz	512	hb.189.gz
512	hb.022.gz	504	hb.064.gz	504	hb.106.gz	504	hb.148.gz	512	hb.190.gz
520	hb.023.gz	512	hb.065.gz	528	hb.107.gz	512	hb.149.gz	512	hb.191.gz
512	hb.024.gz	504	hb.066.gz	520	hb.108.gz	504	hb.150.gz	520	hb.192.gz
512	hb.025.gz	512	hb.067.gz	512	hb.109.gz	504	hb.151.gz	536	hb.193.gz
512	hb.026.gz	520	hb.068.gz	520	hb.110.gz	504	hb.152.gz	520	hb.194.gz
520	hb.027.gz	512	hb.069.gz	512	hb.111.gz	512	hb.153.gz	528	hb.195.gz
512	hb.028.gz	512	hb.070.gz	512	hb.112.gz	512	hb.154.gz	512	hb.196.gz
512	hb.029.gz	520	hb.071.gz	536	hb.113.gz	520	hb.155.gz	528	hb.197.gz
520	hb.030.gz	520	hb.072.gz	448	hb.114.gz	512	hb.156.gz	520	hb.198.gz
520	hb.031.gz	512	hb.073.gz	448	hb.115.gz	512	hb.157.gz	528	hb.199.gz
512	hb.032.gz	528	hb.074.gz	456	hb.116.gz	512	hb.158.gz	512	hb.200.gz
512	hb.033.gz	520	hb.075.gz	448	hb.117.gz	448	hb.159.gz	528	hb.201.gz
512	hb.034.gz	520	hb.076.gz	448	hb.118.gz	448	hb.160.gz	520	hb.202.gz
512	hb.035.gz	520	hb.077.gz	456	hb.119.gz	448	hb.161.gz	520	hb.203.gz
520	hb.036.gz	512	hb.078.gz	456	hb.120.gz	440	hb.162.gz	528	hb.204.gz
520	hb.037.gz	528	hb.079.gz	456	hb.121.gz	440	hb.163.gz	520	hb.205.gz
512	hb.038.gz	512	hb.080.gz	456	hb.122.gz	440	hb.164.gz	528	hb.206.gz
512	hb.039.gz	520	hb.081.gz	448	hb.123.gz	440	hb.165.gz	528	hb.207.gz
520	hb.040.gz	520	hb.082.gz	456	hb.124.gz	448	hb.166.gz	520	hb.208.gz
520	hb.041.gz	520	hb.083.gz	448	hb.125.gz	440	hb.167.gz	520	hb.209.gz
512	hb.042.gz	520	hb.084.gz	448	hb.126.gz	440	hb.168.gz	512	hb.210.gz

MLCC-RELEASE/MLCC-GERMAN/prep: total 14

1	add-headers	1	new-tei-trailer	1	prepare
1	extract	2	prep.perl	1	split.perl
1	final	1	prep2.perl	1	umlauts.perl
3	new-tei-header	1	prep3.perl		

C.3.6 Italian Newspaper Corpus

MLCC-RELEASE/MLCC-ITALIAN: total 19

1	00COPYRIGHT	7	doc.body	1	licence
2	00README	1	doc.tex	1	orig
1	data	4	editdecl.txt	1	prep

MLCC-RELEASE/MLCC-ITALIAN/data: total 4848

200	sole00.tei.gz	192	sole09.tei.gz	168	sole18.tei.gz
208	sole01.tei.gz	184	sole10.tei.gz	208	sole19.tei.gz
176	sole02.tei.gz	184	sole11.tei.gz	168	sole20.tei.gz
208	sole03.tei.gz	208	sole12.tei.gz	192	sole21.tei.gz
176	sole04.tei.gz	184	sole13.tei.gz	152	sole22.tei.gz
208	sole05.tei.gz	208	sole14.tei.gz	208	sole23.tei.gz
160	sole06.tei.gz	200	sole15.tei.gz	168	sole24.tei.gz
200	sole07.tei.gz	176	sole16.tei.gz	136	sole25.tei.gz
200	sole08.tei.gz	176	sole17.tei.gz		

MLCC-RELEASE/MLCC-ITALIAN/licence: total 1

1	00README

MLCC-RELEASE/MLCC-ITALIAN/orig: total 4720

200	sole00.txt.gz	184	sole09.txt.gz	160	sole18.txt.gz
200	sole01.txt.gz	176	sole10.txt.gz	208	sole19.txt.gz
168	sole02.txt.gz	176	sole11.txt.gz	160	sole20.txt.gz
200	sole03.txt.gz	200	sole12.txt.gz	192	sole21.txt.gz
168	sole04.txt.gz	176	sole13.txt.gz	152	sole22.txt.gz
200	sole05.txt.gz	200	sole14.txt.gz	200	sole23.txt.gz
152	sole06.txt.gz	192	sole15.txt.gz	168	sole24.txt.gz
200	sole07.txt.gz	176	sole16.txt.gz	136	sole25.txt.gz
200	sole08.txt.gz	176	sole17.txt.gz		

MLCC-RELEASE/MLCC-ITALIAN/prep: total 34

2	TODO	4	histogram3	1	odd-words
6	conv.perl	4	italian.dtd	8	words.pl
4	histogram	3	new-tei-header		
1	histogram2	1	new-tei-trailer		

C.3.7 Parallel Written Questions Corpus

MLCC-RELEASE/MLCC-JOCWQ: total 70

1	00COPYRIGHT	34	doc.body	1	licence
6	00README	1	doc.tex	12	orig
13	data	1	editdecl.txt	1	prep

MLCC-RELEASE/MLCC-JOCWQ/data: total 24581

59 exp.joc006.93.da.01.tei.gz
63 exp.joc006.93.de.01.tei.gz
56 exp.joc006.93.en.01.tei.gz
61 exp.joc006.93.es.01.tei.gz
61 exp.joc006.93.fr.01.tei.gz
67 exp.joc006.93.gr.01.tei.gz
61 exp.joc006.93.it.01.tei.gz
62 exp.joc006.93.nl.01.tei.gz
61 exp.joc006.93.pt.01.tei.gz
50 exp.joc016.93.da.01.tei.gz
54 exp.joc016.93.de.01.tei.gz
48 exp.joc016.93.en.01.tei.gz
53 exp.joc016.93.es.01.tei.gz
53 exp.joc016.93.fr.01.tei.gz
58 exp.joc016.93.gr.01.tei.gz
52 exp.joc016.93.it.01.tei.gz
53 exp.joc016.93.nl.01.tei.gz
52 exp.joc016.93.pt.01.tei.gz
75 exp.joc032.93.da.01.tei.gz
79 exp.joc032.93.de.01.tei.gz
70 exp.joc032.93.en.01.tei.gz
77 exp.joc032.93.es.01.tei.gz
79 exp.joc032.93.fr.01.tei.gz
85 exp.joc032.93.gr.01.tei.gz
77 exp.joc032.93.it.01.tei.gz
78 exp.joc032.93.nl.01.tei.gz
77 exp.joc032.93.pt.01.tei.gz
76 exp.joc040.93.da.01.tei.gz
82 exp.joc040.93.de.01.tei.gz
72 exp.joc040.93.en.01.tei.gz
79 exp.joc040.93.es.01.tei.gz
80 exp.joc040.93.fr.01.tei.gz
86 exp.joc040.93.gr.01.tei.gz
78 exp.joc040.93.it.01.tei.gz
79 exp.joc040.93.nl.01.tei.gz
79 exp.joc040.93.pt.01.tei.gz
48 exp.joc047.93.da.01.tei.gz
51 exp.joc047.93.de.01.tei.gz
46 exp.joc047.93.en.01.tei.gz
50 exp.joc047.93.es.01.tei.gz
51 exp.joc047.93.fr.01.tei.gz
56 exp.joc047.93.gr.01.tei.gz
50 exp.joc047.93.it.01.tei.gz
51 exp.joc047.93.nl.01.tei.gz
50 exp.joc047.93.pt.01.tei.gz
46 exp.joc051.93.da.01.tei.gz
49 exp.joc051.93.de.01.tei.gz
44 exp.joc051.93.en.01.tei.gz
48 exp.joc051.93.es.01.tei.gz
48 exp.joc051.93.fr.01.tei.gz
53 exp.joc051.93.gr.01.tei.gz
47 exp.joc051.93.it.01.tei.gz
48 exp.joc051.93.nl.01.tei.gz
48 exp.joc051.93.pt.01.tei.gz
62 exp.joc058.93.da.01.tei.gz
67 exp.joc058.93.de.01.tei.gz
59 exp.joc058.93.en.01.tei.gz
65 exp.joc058.93.es.01.tei.gz
66 exp.joc058.93.fr.01.tei.gz
72 exp.joc058.93.gr.01.tei.gz
64 exp.joc058.93.it.01.tei.gz
65 exp.joc058.93.nl.01.tei.gz
65 exp.joc058.93.pt.01.tei.gz
58 exp.joc061.93.da.01.tei.gz
62 exp.joc061.93.de.01.tei.gz
55 exp.joc061.93.en.01.tei.gz
61 exp.joc061.93.es.01.tei.gz
62 exp.joc061.93.fr.01.tei.gz
68 exp.joc061.93.gr.01.tei.gz
60 exp.joc061.93.it.01.tei.gz
61 exp.joc061.93.nl.01.tei.gz
61 exp.joc061.93.pt.01.tei.gz
62 exp.joc065.93.da.01.tei.gz
67 exp.joc065.93.de.01.tei.gz
58 exp.joc065.93.en.01.tei.gz
64 exp.joc065.93.es.01.tei.gz
66 exp.joc065.93.fr.01.tei.gz
72 exp.joc065.93.gr.01.tei.gz
64 exp.joc065.93.it.01.tei.gz
65 exp.joc065.93.nl.01.tei.gz
65 exp.joc065.93.pt.01.tei.gz
42 exp.joc081.93.da.01.tei.gz
46 exp.joc081.93.de.01.tei.gz
40 exp.joc081.93.en.01.tei.gz
44 exp.joc081.93.es.01.tei.gz
44 exp.joc081.93.fr.01.tei.gz
48 exp.joc081.93.gr.01.tei.gz
43 exp.joc081.93.it.01.tei.gz
44 exp.joc081.93.nl.01.tei.gz
44 exp.joc081.93.pt.01.tei.gz
59 exp.joc086.93.da.01.tei.gz
63 exp.joc086.93.de.01.tei.gz
56 exp.joc086.93.en.01.tei.gz
62 exp.joc086.93.es.01.tei.gz
63 exp.joc086.93.fr.01.tei.gz
68 exp.joc086.93.gr.01.tei.gz
61 exp.joc086.93.it.01.tei.gz
62 exp.joc086.93.nl.01.tei.gz
62 exp.joc086.93.pt.01.tei.gz
57 exp.joc090.93.da.01.tei.gz
62 exp.joc090.93.de.01.tei.gz
54 exp.joc090.93.en.01.tei.gz
59 exp.joc090.93.es.01.tei.gz
61 exp.joc090.93.fr.01.tei.gz
65 exp.joc090.93.gr.01.tei.gz
59 exp.joc090.93.it.01.tei.gz
60 exp.joc090.93.nl.01.tei.gz
60 exp.joc090.93.pt.01.tei.gz
60 exp.joc095.93.da.01.tei.gz
64 exp.joc095.93.de.01.tei.gz
56 exp.joc095.93.en.01.tei.gz
62 exp.joc095.93.es.01.tei.gz
62 exp.joc095.93.fr.01.tei.gz
68 exp.joc095.93.gr.01.tei.gz
62 exp.joc095.93.it.01.tei.gz
62 exp.joc095.93.nl.01.tei.gz
62 exp.joc095.93.pt.01.tei.gz
41 exp.joc099.93.da.01.tei.gz
44 exp.joc099.93.de.01.tei.gz
39 exp.joc099.93.en.01.tei.gz
43 exp.joc099.93.es.01.tei.gz
43 exp.joc099.93.fr.01.tei.gz
47 exp.joc099.93.gr.01.tei.gz
43 exp.joc099.93.it.01.tei.gz
43 exp.joc099.93.nl.01.tei.gz
43 exp.joc099.93.pt.01.tei.gz
61 exp.joc101.93.da.01.tei.gz
66 exp.joc101.93.de.01.tei.gz
58 exp.joc101.93.en.01.tei.gz
63 exp.joc101.93.es.01.tei.gz
64 exp.joc101.93.fr.01.tei.gz
70 exp.joc101.93.gr.01.tei.gz
63 exp.joc101.93.it.01.tei.gz
64 exp.joc101.93.nl.01.tei.gz
64 exp.joc101.93.pt.01.tei.gz
69 exp.joc106.93.da.01.tei.gz
75 exp.joc106.93.de.01.tei.gz
66 exp.joc106.93.en.01.tei.gz
73 exp.joc106.93.es.01.tei.gz
74 exp.joc106.93.fr.01.tei.gz
80 exp.joc106.93.gr.01.tei.gz
72 exp.joc106.93.it.01.tei.gz
73 exp.joc106.93.nl.01.tei.gz
73 exp.joc106.93.pt.01.tei.gz
79 exp.joc127.93.da.01.tei.gz
86 exp.joc127.93.de.01.tei.gz
76 exp.joc127.93.en.01.tei.gz
83 exp.joc127.93.es.01.tei.gz
84 exp.joc127.93.fr.01.tei.gz
91 exp.joc127.93.gr.01.tei.gz
82 exp.joc127.93.it.01.tei.gz
83 exp.joc127.93.nl.01.tei.gz
83 exp.joc127.93.pt.01.tei.gz
48 exp.joc132.93.da.01.tei.gz
51 exp.joc132.93.de.01.tei.gz
46 exp.joc132.93.en.01.tei.gz
50 exp.joc132.93.es.01.tei.gz
51 exp.joc132.93.fr.01.tei.gz
56 exp.joc132.93.gr.01.tei.gz
50 exp.joc132.93.it.01.tei.gz
50 exp.joc132.93.nl.01.tei.gz
50 exp.joc132.93.pt.01.tei.gz
48 exp.joc137.93.da.01.tei.gz
51 exp.joc137.93.de.01.tei.gz
45 exp.joc137.93.en.01.tei.gz
49 exp.joc137.93.es.01.tei.gz
50 exp.joc137.93.fr.01.tei.gz
54 exp.joc137.93.gr.01.tei.gz
49 exp.joc137.93.it.01.tei.gz
50 exp.joc137.93.nl.01.tei.gz
49 exp.joc137.93.pt.01.tei.gz
65 exp.joc141.93.da.01.tei.gz
70 exp.joc141.93.de.01.tei.gz
62 exp.joc141.93.en.01.tei.gz
68 exp.joc141.93.es.01.tei.gz
69 exp.joc141.93.fr.01.tei.gz
75 exp.joc141.93.gr.01.tei.gz
68 exp.joc141.93.it.01.tei.gz
69 exp.joc141.93.nl.01.tei.gz
69 exp.joc141.93.pt.01.tei.gz
60 exp.joc145.93.da.01.tei.gz
65 exp.joc145.93.de.01.tei.gz
57 exp.joc145.93.en.01.tei.gz
64 exp.joc145.93.es.01.tei.gz
64 exp.joc145.93.fr.01.tei.gz
70 exp.joc145.93.gr.01.tei.gz
63 exp.joc145.93.it.01.tei.gz
63 exp.joc145.93.nl.01.tei.gz
63 exp.joc145.93.pt.01.tei.gz
57 exp.joc155.93.da.01.tei.gz
61 exp.joc155.93.de.01.tei.gz
54 exp.joc155.93.en.01.tei.gz
59 exp.joc155.93.es.01.tei.gz
60 exp.joc155.93.fr.01.tei.gz
66 exp.joc155.93.gr.01.tei.gz
59 exp.joc155.93.it.01.tei.gz
59 exp.joc155.93.nl.01.tei.gz
59 exp.joc155.93.pt.01.tei.gz
41 exp.joc162.93.da.01.tei.gz
44 exp.joc162.93.de.01.tei.gz
39 exp.joc162.93.en.01.tei.gz
43 exp.joc162.93.es.01.tei.gz
44 exp.joc162.93.fr.01.tei.gz
47 exp.joc162.93.gr.01.tei.gz
43 exp.joc162.93.it.01.tei.gz
43 exp.joc162.93.nl.01.tei.gz
43 exp.joc162.93.pt.01.tei.gz
82 exp.joc185.93.da.01.tei.gz
88 exp.joc185.93.de.01.tei.gz
78 exp.joc185.93.en.01.tei.gz
85 exp.joc185.93.es.01.tei.gz
87 exp.joc185.93.fr.01.tei.gz
94 exp.joc185.93.gr.01.tei.gz
86 exp.joc185.93.it.01.tei.gz
86 exp.joc185.93.nl.01.tei.gz
85 exp.joc185.93.pt.01.tei.gz
112 exp.joc195.93.da.01.tei.gz
120 exp.joc195.93.de.01.tei.gz
93 exp.joc195.93.en.01.tei.gz
112 exp.joc195.93.es.01.tei.gz
120 exp.joc195.93.fr.01.tei.gz
128 exp.joc195.93.gr.01.tei.gz
112 exp.joc195.93.it.01.tei.gz
112 exp.joc195.93.nl.01.tei.gz
112 exp.joc195.93.pt.01.tei.gz
56 exp.joc202.93.da.01.tei.gz
60 exp.joc202.93.de.01.tei.gz
53 exp.joc202.93.en.01.tei.gz
58 exp.joc202.93.es.01.tei.gz
59 exp.joc202.93.fr.01.tei.gz
64 exp.joc202.93.gr.01.tei.gz
58 exp.joc202.93.it.01.tei.gz
59 exp.joc202.93.nl.01.tei.gz
58 exp.joc202.93.pt.01.tei.gz
71 exp.joc207.93.da.01.tei.gz
77 exp.joc207.93.de.01.tei.gz
68 exp.joc207.93.en.01.tei.gz
74 exp.joc207.93.es.01.tei.gz
77 exp.joc207.93.fr.01.tei.gz
82 exp.joc207.93.gr.01.tei.gz
75 exp.joc207.93.it.01.tei.gz
75 exp.joc207.93.nl.01.tei.gz
74 exp.joc207.93.pt.01.tei.gz
36 exp.joc241.93.da.01.tei.gz
40 exp.joc241.93.de.01.tei.gz
35 exp.joc241.93.en.01.tei.gz
39 exp.joc241.93.es.01.tei.gz
39 exp.joc241.93.fr.01.tei.gz
42 exp.joc241.93.gr.01.tei.gz
38 exp.joc241.93.it.01.tei.gz
38 exp.joc241.93.nl.01.tei.gz
38 exp.joc241.93.pt.01.tei.gz
95 exp.joc258.93.da.01.tei.gz
112 exp.joc258.93.de.01.tei.gz
90 exp.joc258.93.en.01.tei.gz
112 exp.joc258.93.es.01.tei.gz
112 exp.joc258.93.fr.01.tei.gz
120 exp.joc258.93.gr.01.tei.gz
112 exp.joc258.93.it.01.tei.gz
112 exp.joc258.93.nl.01.tei.gz
112 exp.joc258.93.pt.01.tei.gz
96 exp.joc264.93.da.01.tei.gz
112 exp.joc264.93.de.01.tei.gz
90 exp.joc264.93.en.01.tei.gz
112 exp.joc264.93.es.01.tei.gz
112 exp.joc264.93.fr.01.tei.gz
120 exp.joc264.93.gr.01.tei.gz
112 exp.joc264.93.it.01.tei.gz
112 exp.joc264.93.nl.01.tei.gz
112 exp.joc264.93.pt.01.tei.gz
112 exp.joc280.93.da.01.tei.gz
112 exp.joc280.93.de.01.tei.gz
92 exp.joc280.93.en.01.tei.gz
112 exp.joc280.93.es.01.tei.gz
112 exp.joc280.93.fr.01.tei.gz
120 exp.joc280.93.gr.01.tei.gz
112 exp.joc280.93.it.01.tei.gz
112 exp.joc280.93.nl.01.tei.gz
112 exp.joc280.93.pt.01.tei.gz
81 exp.joc283.93.da.01.tei.gz
87 exp.joc283.93.de.01.tei.gz
77 exp.joc283.93.en.01.tei.gz
85 exp.joc283.93.es.01.tei.gz
86 exp.joc283.93.fr.01.tei.gz
93 exp.joc283.93.gr.01.tei.gz
84 exp.joc283.93.it.01.tei.gz
85 exp.joc283.93.nl.01.tei.gz
83 exp.joc283.93.pt.01.tei.gz
57 exp.joc288.93.da.01.tei.gz
61 exp.joc288.93.de.01.tei.gz
54 exp.joc288.93.en.01.tei.gz
59 exp.joc288.93.es.01.tei.gz
61 exp.joc288.93.fr.01.tei.gz
66 exp.joc288.93.gr.01.tei.gz
58 exp.joc288.93.it.01.tei.gz
60 exp.joc288.93.nl.01.tei.gz
59 exp.joc288.93.pt.01.tei.gz
112 exp.joc292.93.da.01.tei.gz
120 exp.joc292.93.de.01.tei.gz
96 exp.joc292.93.en.01.tei.gz
120 exp.joc292.93.es.01.tei.gz
120 exp.joc292.93.fr.01.tei.gz
128 exp.joc292.93.gr.01.tei.gz
120 exp.joc292.93.it.01.tei.gz
120 exp.joc292.93.nl.01.tei.gz
120 exp.joc292.93.pt.01.tei.gz
60 exp.joc297.93.da.01.tei.gz
65 exp.joc297.93.de.01.tei.gz
57 exp.joc297.93.en.01.tei.gz
63 exp.joc297.93.es.01.tei.gz
64 exp.joc297.93.fr.01.tei.gz
69 exp.joc297.93.gr.01.tei.gz
62 exp.joc297.93.it.01.tei.gz
63 exp.joc297.93.nl.01.tei.gz
63 exp.joc297.93.pt.01.tei.gz
68 exp.joc301.93.da.01.tei.gz
74 exp.joc301.93.de.01.tei.gz
65 exp.joc301.93.en.01.tei.gz
71 exp.joc301.93.es.01.tei.gz
73 exp.joc301.93.fr.01.tei.gz
78 exp.joc301.93.gr.01.tei.gz
71 exp.joc301.93.it.01.tei.gz
72 exp.joc301.93.nl.01.tei.gz
72 exp.joc301.93.pt.01.tei.gz
67 exp.joc320.93.da.01.tei.gz
72 exp.joc320.93.de.01.tei.gz
64 exp.joc320.93.en.01.tei.gz
71 exp.joc320.93.es.01.tei.gz
72 exp.joc320.93.fr.01.tei.gz
78 exp.joc320.93.gr.01.tei.gz
69 exp.joc320.93.it.01.tei.gz
71 exp.joc320.93.nl.01.tei.gz
70 exp.joc320.93.pt.01.tei.gz
60 exp.joc327.93.da.01.tei.gz
64 exp.joc327.93.de.01.tei.gz
57 exp.joc327.93.en.01.tei.gz
62 exp.joc327.93.es.01.tei.gz
63 exp.joc327.93.fr.01.tei.gz
69 exp.joc327.93.gr.01.tei.gz
62 exp.joc327.93.it.01.tei.gz
63 exp.joc327.93.nl.01.tei.gz
62 exp.joc327.93.pt.01.tei.gz
47 exp.joc333.93.da.01.tei.gz
51 exp.joc333.93.de.01.tei.gz
45 exp.joc333.93.en.01.tei.gz
49 exp.joc333.93.es.01.tei.gz
51 exp.joc333.93.fr.01.tei.gz
55 exp.joc333.93.gr.01.tei.gz
49 exp.joc333.93.it.01.tei.gz
49 exp.joc333.93.nl.01.tei.gz
49 exp.joc333.93.pt.01.tei.gz
69 exp.joc350.93.da.01.tei.gz
75 exp.joc350.93.de.01.tei.gz
66 exp.joc350.93.en.01.tei.gz
72 exp.joc350.93.es.01.tei.gz
75 exp.joc350.93.fr.01.tei.gz
79 exp.joc350.93.gr.01.tei.gz
71 exp.joc350.93.it.01.tei.gz
72 exp.joc350.93.nl.01.tei.gz
72 exp.joc350.93.pt.01.tei.gz

MLCC-RELEASE/MLCC-JOCWQ/licence: total 12

1	00README	6	non-disclosure-agr-opoce.tex
5	agr-ed-multext-opoce.tex		

MLCC-RELEASE/MLCC-JOCWQ/orig: total 26632

1	README
64	exp.joc006.93.da.01.gz
68	exp.joc006.93.de.01.gz
60	exp.joc006.93.en.01.gz
66	exp.joc006.93.es.01.gz
68	exp.joc006.93.fr.01.gz
74	exp.joc006.93.gr.01.gz
66	exp.joc006.93.it.01.gz
67	exp.joc006.93.nl.01.gz
67	exp.joc006.93.pt.01.gz
54	exp.joc016.93.da.01.gz
58	exp.joc016.93.de.01.gz
52	exp.joc016.93.en.01.gz
57	exp.joc016.93.es.01.gz
58	exp.joc016.93.fr.01.gz
64	exp.joc016.93.gr.01.gz
56	exp.joc016.93.it.01.gz
57	exp.joc016.93.nl.01.gz
57	exp.joc016.93.pt.01.gz
82	exp.joc032.93.da.01.gz
86	exp.joc032.93.de.01.gz
76	exp.joc032.93.en.01.gz
84	exp.joc032.93.es.01.gz
87	exp.joc032.93.fr.01.gz
95	exp.joc032.93.gr.01.gz
84	exp.joc032.93.it.01.gz
84	exp.joc032.93.nl.01.gz
85	exp.joc032.93.pt.01.gz
83	exp.joc040.93.da.01.gz
89	exp.joc040.93.de.01.gz
80	exp.joc040.93.en.01.gz
87	exp.joc040.93.es.01.gz
90	exp.joc040.93.fr.01.gz
112	exp.joc040.93.gr.01.gz
86	exp.joc040.93.it.01.gz
86	exp.joc040.93.nl.01.gz
88	exp.joc040.93.pt.01.gz
52	exp.joc047.93.da.01.gz
55	exp.joc047.93.de.01.gz
50	exp.joc047.93.en.01.gz
54	exp.joc047.93.es.01.gz
56	exp.joc047.93.fr.01.gz
62	exp.joc047.93.gr.01.gz
54	exp.joc047.93.it.01.gz
54	exp.joc047.93.nl.01.gz
55	exp.joc047.93.pt.01.gz
50	exp.joc051.93.da.01.gz
54	exp.joc051.93.de.01.gz
47	exp.joc051.93.en.01.gz
52	exp.joc051.93.es.01.gz
53	exp.joc051.93.fr.01.gz
58	exp.joc051.93.gr.01.gz
51	exp.joc051.93.it.01.gz
52	exp.joc051.93.nl.01.gz
53	exp.joc051.93.pt.01.gz
67	exp.joc058.93.da.01.gz
73	exp.joc058.93.de.01.gz
64	exp.joc058.93.en.01.gz
70	exp.joc058.93.es.01.gz
72	exp.joc058.93.fr.01.gz
79	exp.joc058.93.gr.01.gz
69	exp.joc058.93.it.01.gz
70	exp.joc058.93.nl.01.gz
71	exp.joc058.93.pt.01.gz
63	exp.joc061.93.da.01.gz
68	exp.joc061.93.de.01.gz
60	exp.joc061.93.en.01.gz
66	exp.joc061.93.es.01.gz
68	exp.joc061.93.fr.01.gz
75	exp.joc061.93.gr.01.gz
65	exp.joc061.93.it.01.gz
66	exp.joc061.93.nl.01.gz
67	exp.joc061.93.pt.01.gz
67	exp.joc065.93.da.01.gz
72	exp.joc065.93.de.01.gz
63	exp.joc065.93.en.01.gz
70	exp.joc065.93.es.01.gz
72	exp.joc065.93.fr.01.gz
79	exp.joc065.93.gr.01.gz
69	exp.joc065.93.it.01.gz
70	exp.joc065.93.nl.01.gz
71	exp.joc065.93.pt.01.gz
45	exp.joc081.93.da.01.gz
49	exp.joc081.93.de.01.gz
44	exp.joc081.93.en.01.gz
47	exp.joc081.93.es.01.gz
49	exp.joc081.93.fr.01.gz
53	exp.joc081.93.gr.01.gz
47	exp.joc081.93.it.01.gz
47	exp.joc081.93.nl.01.gz
48	exp.joc081.93.pt.01.gz
64	exp.joc086.93.da.01.gz
69	exp.joc086.93.de.01.gz
61	exp.joc086.93.en.01.gz
68	exp.joc086.93.es.01.gz
69	exp.joc086.93.fr.01.gz
76	exp.joc086.93.gr.01.gz
67	exp.joc086.93.it.01.gz
67	exp.joc086.93.nl.01.gz
69	exp.joc086.93.pt.01.gz
62	exp.joc090.93.da.01.gz
67	exp.joc090.93.de.01.gz
59	exp.joc090.93.en.01.gz
65	exp.joc090.93.es.01.gz
67	exp.joc090.93.fr.01.gz
72	exp.joc090.93.gr.01.gz
64	exp.joc090.93.it.01.gz
64	exp.joc090.93.nl.01.gz
66	exp.joc090.93.pt.01.gz
64	exp.joc095.93.da.01.gz
69	exp.joc095.93.de.01.gz
60	exp.joc095.93.en.01.gz
67	exp.joc095.93.es.01.gz
68	exp.joc095.93.fr.01.gz
75	exp.joc095.93.gr.01.gz
66	exp.joc095.93.it.01.gz
67	exp.joc095.93.nl.01.gz
68	exp.joc095.93.pt.01.gz
44	exp.joc099.93.da.01.gz
48	exp.joc099.93.de.01.gz
43	exp.joc099.93.en.01.gz
47	exp.joc099.93.es.01.gz
48	exp.joc099.93.fr.01.gz
52	exp.joc099.93.gr.01.gz
46	exp.joc099.93.it.01.gz
46	exp.joc099.93.nl.01.gz
47	exp.joc099.93.pt.01.gz
67	exp.joc101.93.da.01.gz
71	exp.joc101.93.de.01.gz
63	exp.joc101.93.en.01.gz
69	exp.joc101.93.es.01.gz
71	exp.joc101.93.fr.01.gz
77	exp.joc101.93.gr.01.gz
68	exp.joc101.93.it.01.gz
69	exp.joc101.93.nl.01.gz
70	exp.joc101.93.pt.01.gz
76	exp.joc106.93.da.01.gz
82	exp.joc106.93.de.01.gz
72	exp.joc106.93.en.01.gz
80	exp.joc106.93.es.01.gz
82	exp.joc106.93.fr.01.gz
89	exp.joc106.93.gr.01.gz
79	exp.joc106.93.it.01.gz
80	exp.joc106.93.nl.01.gz
81	exp.joc106.93.pt.01.gz
86	exp.joc127.93.da.01.gz
94	exp.joc127.93.de.01.gz
83	exp.joc127.93.en.01.gz
92	exp.joc127.93.es.01.gz
93	exp.joc127.93.fr.01.gz
112	exp.joc127.93.gr.01.gz
89	exp.joc127.93.it.01.gz
90	exp.joc127.93.nl.01.gz
93	exp.joc127.93.pt.01.gz
52	exp.joc132.93.da.01.gz
56	exp.joc132.93.de.01.gz
50	exp.joc132.93.en.01.gz
55	exp.joc132.93.es.01.gz
56	exp.joc132.93.fr.01.gz
61	exp.joc132.93.gr.01.gz
54	exp.joc132.93.it.01.gz
54	exp.joc132.93.nl.01.gz
55	exp.joc132.93.pt.01.gz
52	exp.joc137.93.da.01.gz
55	exp.joc137.93.de.01.gz
49	exp.joc137.93.en.01.gz
54	exp.joc137.93.es.01.gz
55	exp.joc137.93.fr.01.gz
60	exp.joc137.93.gr.01.gz
53	exp.joc137.93.it.01.gz
53	exp.joc137.93.nl.01.gz
54	exp.joc137.93.pt.01.gz
71	exp.joc141.93.da.01.gz
76	exp.joc141.93.de.01.gz
68	exp.joc141.93.en.01.gz
74	exp.joc141.93.es.01.gz
76	exp.joc141.93.fr.01.gz
83	exp.joc141.93.gr.01.gz
74	exp.joc141.93.it.01.gz
74	exp.joc141.93.nl.01.gz
76	exp.joc141.93.pt.01.gz
67	exp.joc145.93.da.01.gz
72	exp.joc145.93.de.01.gz
64	exp.joc145.93.en.01.gz
71	exp.joc145.93.es.01.gz
72	exp.joc145.93.fr.01.gz
78	exp.joc145.93.gr.01.gz
70	exp.joc145.93.it.01.gz
69	exp.joc145.93.nl.01.gz
70	exp.joc145.93.pt.01.gz
62	exp.joc155.93.da.01.gz
66	exp.joc155.93.de.01.gz
58	exp.joc155.93.en.01.gz
64	exp.joc155.93.es.01.gz
66	exp.joc155.93.fr.01.gz
73	exp.joc155.93.gr.01.gz
64	exp.joc155.93.it.01.gz
64	exp.joc155.93.nl.01.gz
65	exp.joc155.93.pt.01.gz
44	exp.joc162.93.da.01.gz
47	exp.joc162.93.de.01.gz
42	exp.joc162.93.en.01.gz
47	exp.joc162.93.es.01.gz
48	exp.joc162.93.fr.01.gz
51	exp.joc162.93.gr.01.gz
46	exp.joc162.93.it.01.gz
46	exp.joc162.93.nl.01.gz
46	exp.joc162.93.pt.01.gz
89	exp.joc185.93.da.01.gz
96	exp.joc185.93.de.01.gz
84	exp.joc185.93.en.01.gz
93	exp.joc185.93.es.01.gz
112	exp.joc185.93.fr.01.gz
112	exp.joc185.93.gr.01.gz
94	exp.joc185.93.it.01.gz
92	exp.joc185.93.nl.01.gz
94	exp.joc185.93.pt.01.gz
120	exp.joc195.93.da.01.gz
128	exp.joc195.93.de.01.gz
112	exp.joc195.93.en.01.gz
128	exp.joc195.93.es.01.gz
128	exp.joc195.93.fr.01.gz
136	exp.joc195.93.gr.01.gz
120	exp.joc195.93.it.01.gz
120	exp.joc195.93.nl.01.gz
128	exp.joc195.93.pt.01.gz
60	exp.joc202.93.da.01.gz
64	exp.joc202.93.de.01.gz
57	exp.joc202.93.en.01.gz
63	exp.joc202.93.es.01.gz
65	exp.joc202.93.fr.01.gz
70	exp.joc202.93.gr.01.gz
63	exp.joc202.93.it.01.gz
63	exp.joc202.93.nl.01.gz
64	exp.joc202.93.pt.01.gz
77	exp.joc207.93.da.01.gz
83	exp.joc207.93.de.01.gz
73	exp.joc207.93.en.01.gz
81	exp.joc207.93.es.01.gz
84	exp.joc207.93.fr.01.gz
90	exp.joc207.93.gr.01.gz
81	exp.joc207.93.it.01.gz
80	exp.joc207.93.nl.01.gz
82	exp.joc207.93.pt.01.gz
39	exp.joc241.93.da.01.gz
43	exp.joc241.93.de.01.gz
38	exp.joc241.93.en.01.gz
42	exp.joc241.93.es.01.gz
43	exp.joc241.93.fr.01.gz
45	exp.joc241.93.gr.01.gz
41	exp.joc241.93.it.01.gz
40	exp.joc241.93.nl.01.gz
42	exp.joc241.93.pt.01.gz
112	exp.joc258.93.da.01.gz
120	exp.joc258.93.de.01.gz
112	exp.joc258.93.en.01.gz
120	exp.joc258.93.es.01.gz
120	exp.joc258.93.fr.01.gz
128	exp.joc258.93.gr.01.gz
120	exp.joc258.93.it.01.gz
120	exp.joc258.93.nl.01.gz
120	exp.joc258.93.pt.01.gz
120	exp.joc264.93.da.01.gz
120	exp.joc264.93.de.01.gz
112	exp.joc264.93.en.01.gz
120	exp.joc264.93.es.01.gz
120	exp.joc264.93.fr.01.gz
128	exp.joc264.93.gr.01.gz
120	exp.joc264.93.it.01.gz
120	exp.joc264.93.nl.01.gz
120	exp.joc264.93.pt.01.gz
120	exp.joc280.93.da.01.gz
128	exp.joc280.93.de.01.gz
112	exp.joc280.93.en.01.gz
120	exp.joc280.93.es.01.gz
128	exp.joc280.93.fr.01.gz
136	exp.joc280.93.gr.01.gz
120	exp.joc280.93.it.01.gz
120	exp.joc280.93.nl.01.gz
120	exp.joc280.93.pt.01.gz
84	exp.joc283.93.da.01.gz
91	exp.joc283.93.de.01.gz
79	exp.joc283.93.en.01.gz
89	exp.joc283.93.es.01.gz
91	exp.joc283.93.fr.01.gz
96	exp.joc283.93.gr.01.gz
87	exp.joc283.93.it.01.gz
88	exp.joc283.93.nl.01.gz
88	exp.joc283.93.pt.01.gz
59	exp.joc288.93.da.01.gz
63	exp.joc288.93.de.01.gz
56	exp.joc288.93.en.01.gz
62	exp.joc288.93.es.01.gz
64	exp.joc288.93.fr.01.gz
68	exp.joc288.93.gr.01.gz
60	exp.joc288.93.it.01.gz
62	exp.joc288.93.nl.01.gz
62	exp.joc288.93.pt.01.gz
120	exp.joc292.93.da.01.gz
128	exp.joc292.93.de.01.gz
112	exp.joc292.93.en.01.gz
120	exp.joc292.93.es.01.gz
128	exp.joc292.93.fr.01.gz
136	exp.joc292.93.gr.01.gz
120	exp.joc292.93.it.01.gz
120	exp.joc292.93.nl.01.gz
120	exp.joc292.93.pt.01.gz
63	exp.joc297.93.da.01.gz
68	exp.joc297.93.de.01.gz
59	exp.joc297.93.en.01.gz
66	exp.joc297.93.es.01.gz
67	exp.joc297.93.fr.01.gz
71	exp.joc297.93.gr.01.gz
65	exp.joc297.93.it.01.gz
65	exp.joc297.93.nl.01.gz
66	exp.joc297.93.pt.01.gz
71	exp.joc301.93.da.01.gz
77	exp.joc301.93.de.01.gz
67	exp.joc301.93.en.01.gz
75	exp.joc301.93.es.01.gz
77	exp.joc301.93.fr.01.gz
81	exp.joc301.93.gr.01.gz
74	exp.joc301.93.it.01.gz
74	exp.joc301.93.nl.01.gz
76	exp.joc301.93.pt.01.gz
70	exp.joc320.93.da.01.gz
75	exp.joc320.93.de.01.gz
66	exp.joc320.93.en.01.gz
74	exp.joc320.93.es.01.gz
76	exp.joc320.93.fr.01.gz
80	exp.joc320.93.gr.01.gz
72	exp.joc320.93.it.01.gz
73	exp.joc320.93.nl.01.gz
74	exp.joc320.93.pt.01.gz
62	exp.joc327.93.da.01.gz
67	exp.joc327.93.de.01.gz
59	exp.joc327.93.en.01.gz
65	exp.joc327.93.es.01.gz
67	exp.joc327.93.fr.01.gz
71	exp.joc327.93.gr.01.gz
64	exp.joc327.93.it.01.gz
65	exp.joc327.93.nl.01.gz
65	exp.joc327.93.pt.01.gz
49	exp.joc333.93.da.01.gz
53	exp.joc333.93.de.01.gz
46	exp.joc333.93.en.01.gz
51	exp.joc333.93.es.01.gz
53	exp.joc333.93.fr.01.gz
56	exp.joc333.93.gr.01.gz
50	exp.joc333.93.it.01.gz
51	exp.joc333.93.nl.01.gz
51	exp.joc333.93.pt.01.gz
72	exp.joc350.93.da.01.gz
78	exp.joc350.93.de.01.gz
69	exp.joc350.93.en.01.gz
75	exp.joc350.93.es.01.gz
80	exp.joc350.93.fr.01.gz
82	exp.joc350.93.gr.01.gz
74	exp.joc350.93.it.01.gz
75	exp.joc350.93.nl.01.gz
76	exp.joc350.93.pt.01.gz

MLCC-RELEASE/MLCC-JOCWQ/prep: total 90

5	abbrs.pl	1	missing-seps-between-numbers.pl
1	check-p.perl	3	new-tei-header
1	count.pl	1	new-tei-trailer
1	count2.pl	1	numpar.perl
2	generalise.pl	11	prep.perl
2	get_struct.perl	1	script
1	greek-chars.el	1	spin.sgm
3	make-header.perl	2	tei.header.template
52	missing-seps-between-numbers	1	tei.trailer.template

C.3.8 Spanish Newspaper Corpus

MLCC-RELEASE/MLCC-SPANISH: total 18

1	00COPYRIGHT	5	doc.body	1	licence
1	00README	1	doc.tex	1	orig
2	data	5	editdecl.txt	1	prep

MLCC-RELEASE/MLCC-SPANISH/data: total 22944

672	expan1aa.sgm.gz	712	expan1al.sgm.gz	720	expan1aw.sgm.gz
712	expan1ab.sgm.gz	720	expan1am.sgm.gz	712	expan1ax.sgm.gz
704	expan1ac.sgm.gz	696	expan1an.sgm.gz	704	expan1ay.sgm.gz
712	expan1ad.sgm.gz	720	expan1ao.sgm.gz	704	expan1az.sgm.gz
712	expan1ae.sgm.gz	712	expan1ap.sgm.gz	720	expan1ba.sgm.gz
736	expan1af.sgm.gz	720	expan1aq.sgm.gz	736	expan1bb.sgm.gz
712	expan1ag.sgm.gz	704	expan1ar.sgm.gz	704	expan1bc.sgm.gz
680	expan1ah.sgm.gz	704	expan1as.sgm.gz	736	expan1bd.sgm.gz
696	expan1ai.sgm.gz	688	expan1at.sgm.gz	744	expan1be.sgm.gz
696	expan1aj.sgm.gz	712	expan1au.sgm.gz	760	expan1bf.sgm.gz
704	expan1ak.sgm.gz	696	expan1av.sgm.gz	184	expan1bg.sgm.gz

MLCC-RELEASE/MLCC-SPANISH/licence: total 6

6	non-disclosure-agr-exp-2.tex

MLCC-RELEASE/MLCC-SPANISH/orig: total 22864

672	expan1aa.gz	712	expan1al.gz	720	expan1aw.gz
712	expan1ab.gz	720	expan1am.gz	712	expan1ax.gz
696	expan1ac.gz	688	expan1an.gz	704	expan1ay.gz
712	expan1ad.gz	712	expan1ao.gz	704	expan1az.gz
712	expan1ae.gz	704	expan1ap.gz	712	expan1ba.gz
736	expan1af.gz	720	expan1aq.gz	728	expan1bb.gz
712	expan1ag.gz	704	expan1ar.gz	696	expan1bc.gz
680	expan1ah.gz	704	expan1as.gz	736	expan1bd.gz
688	expan1ai.gz	688	expan1at.gz	744	expan1be.gz
696	expan1aj.gz	704	expan1au.gz	760	expan1bf.gz
696	expan1ak.gz	696	expan1av.gz	184	expan1bg.gz

MLCC-RELEASE/MLCC-SPANISH/prep: total 28

1	add-headers	1	new-tei-trailer	2	script
2	clean.perl	1	prepA.perl	3	script2
4	clean2.perl	4	prepB.perl	4	script3
1	final	1	prepC.perl		
3	new-tei-header	1	replace.perl		