[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

MULTILINGUAL CORPORA FOR COOPERATION

Susan Armstrong[1], Masja Kempen[2], David McKelvie[2], Dominique Petitpierre[1], Reinhard Rapp[3] and Henry S. Thompson[2]
[1]ISSCO/ETI, University of Geneva
[2]HCRC, Edinburgh University
[3]FASK, University of Mainz

ABSTRACT

MLCC was a corpus, acquisition project funded by the EC Telematics program.The aim was to collect a set of texts representing a substantial improvement in range, quantity and quality of corpus material available. Two sub-corpora have been defined to help meet the needs for multilingual data consisting of a comparable set of texts in six languages and a parallel set of data in 9 languages. The comparable text collection includes financial newspaper articles from the early '90s. The parallel data is taken from the Official Journal of the European Commission, sub-series Written Questions to Parliament and from the Proceedings of the European Parliament. The data has been converted to an SGML, TEI-conformant mark-up and is distributed by ELRA.

1 INTRODUCTION

We report here on the corpus data acquired and prepared under the "MLCC" project (Multilingual Corpora for Cooperation) from 1994 to 1995. This project was elaborated by LTG, Edinburgh and ISSCO, Geneva with coordination by CNR, Pisa within the framework of the LRE (Linguistic Research and Engineering) program (Armstrong, et al., 1995)[1]. 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 MLCC corpus is intended to answer the needs of the linguistic research community for comparable data in a wide range of languages. The polylingual document collection, consisting of similar documents in six languages, provides an important addition to monolingual collections (such as the CD-ROM of the European Corpus Initiative, cf. Armstrong-Warwick, et al., 1994) and assures that researchers from different countries can carry out similar studies in their own language. This collection, with obvious connections to the TREC/Tipster materials used in the United States, enables comparability in research on an international scale. The polylingual document collection provides the basis for comparable evaluation programs in Information Retrieval and NLP for different European languages. The parallel data consisting of translated data in nine languages provides a wealth of material for translation studies and evaluation of technology development across different language pairs.

In the following sections we first give a brief overview of the corpora and discuss the problems associated with obtaining the data and negotiating distribution rights. Each of the collections are then described in more detail, followed by a discussion of the SGML markup.

1.1 Overview of the corpora

The MLCC text corpus has two main components - one set to allow comparable studies to be carried out in different languages and another 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 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, Portuguese 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, ca. 1,1 million words per language
- Official Journal of the European Commission, Annex: Debates of the European Parliament 1992-1994, 5 to 8 million words per language

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 feasibility of acquiring and preparing such data within the constraints of the project. This corpus will by no means entirely satisfy the need for very large amounts of varied resources in many languages, though it does represent a first step.

The data presented here offers a wealth of opportunities for evaluating NLP methods and systems across a range of European languages. In order to make the MLCC corpus available to a wide audience, licence agreements were set up with the data providers. With the foundation of the European Linguistic Resource Association (ELRA), these agreements have all been re-negotiated in favour of ELRA/ELDA. By now, both the comparable and the parallel corpora can be obtained from ELRA (http://www.icp.grenet.fr/ELRA/home.html). The cost depends on whether the corpora are used for academic or commercial purposes and is lower for members of ELRA.[2]

1.2 Corpus acquisition

Although there is a vast amount of electronic data potentially available, the acquisition process is by no means

[1] The final report of the MLCC project is available on the web at http://www.ltg.ed.ac.uk/~dmck.

[2] From the outset, it was the expressed intention of the MLCC project to make all corpora available at a nominal charge. Unfortunately, pricing policies were not decided by the project members.

straightforward (Armstrong-Warwick, 1994) We can identify a number of steps in this process, each of which present new obstacles to be overcome:

- identify potential collections of electronic data
- contact the relevant organizations and identify the authority for acquisition and distribution rights
- investigate the availability of electronic versions of the data and the technical problems to be overcome
- negotiate with the appropriate authorities
- acquire the data from the technical services
- decode the original formatting
- clean and mark the data
- establish adequate distribution mechanisms
- assure that licence agreements are completed

Each of these steps requires at least a great deal of patience and can imply long delays before any data is physically acquired, let alone cleaned and marked. Large collections of printed data are often no longer available in electronic form and if they do exist it is not always clear where they are being held. In order to negotiate distribution rights, the proper authority must be identified. In many institutions it is not clear which office actually has that role. Once the relevant individuals have been identified, they must be convinced that this is a 'good' and 'safe' endeavor. The lack of public awareness of the need for data in NLP research, compounded by the lack of a proper legal basis to protect against the misuse of electronic data implies that negotiations can indeed take a long time. In addition, it is often difficult to explain to the data holder the idea of texts measured in megabytes for NLP research, rather than in terms of content and hence commercial value.

Once the provider is convinced, legal services must be consulted and then there is still a great deal of work to physically acquire and prepare the data. The technical services where the data is stored is often far removed from the office with the authority to release the data. Simply accessing the data can be difficult due to practical matters such as locating the data and technical issues such as antiquated storage media and formats. And once the material is acquired, it is sometimes encoded in an undocumented and cryptic typesetting language whose decoding can require far more time than expected and would be better done by specialists from the secret services. Examples of how much work can be required to clean and mark the data are documented in Liberman (1989) and McKelvie & Thompson (1994). Another example are the parliamentary debates presented below.

2 Multilingual Parallel Corpus

This part of the project was directed at collecting parallel texts in the 9 European Union official languages (as of 1993), and consists of corpora from the Office of Publications of the European Community (OPOCE). 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. The collection of the Parliamentary Debates, however, proved somewhat more difficult, as discussed below.

2.1 Debates of the European Parliament

The first parallel corpus in 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 meetings. The texts also contain copies of written input provided to the meetings 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. The text 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 comprise the Debates of the European Parliament from January 1992 to July 1994. This sub-corpus contains some 5 to 8 million words per language.

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 content ranges from technical discussions to 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.

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 as a basis for larger translation studies.[3] Such a growing collection

[3] 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

could also serve general language studies in important areas such as language coverage and vocabulary growth.

2.1.1 Data acquisition

Unlike the Written Questions series described in the next section, the acquisition and subsequent preparation of the Debates required considerable time and effort. The physical acquisition alone proceeded in phases over an entire year.[4] 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 Bernoulli diskpacks and 100 diskettes. An overview is given in Table 1 in terms of languages, amount, and formats.

Lang	**IBM-tapes** (MOPAS 1)	**IBM-tapes** (ASCII 1)	**Bernoulli** (MOPAS 2)	**Diskettes** (ASCII 2)
DA	1.7		3.4	1.7
NL	3.4		2.5	
GB	2.5	0.8	4.2	
FR		2.0	2.5	
DE			1.0	7.0
GR			1.9	7.0
IT		5.0	1.9	
PT	0.8	3.4	3.4	
SP	1.7		3.4	

Table 1: Overview of data from the European Parliament (figures are in million words).

For exploratory data conversion work, a first sample tape of data was acquired. The purpose of this sample tape was 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[5] also required the acquisition and installation of a Bernoulli tape reader in order to access the data. The last set of data comprised 100 diskettes from four independent print shops.

2.1.2 Data Preparation

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

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

2.1.3 Reading the Data

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 Spare 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.

Bernoulli disk packs:

In order to read the Bernoulli disk packs, the Parliamentary Printing offices lent ISSCO a Bernoulli 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 Spare station.

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 Spare station.

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 and 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.

2.1.4 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 format

series should now be considerably easier since, as of July 1995, the data preparation process has been modernized.

[4] With considerable help from Mr. Brogard, Director of the European Parliament Printing Services, we were able to locate and obtain the full set of the Debates for the period.

[5] A Bernoulli disk pack can contain up to 90 Mb of data and is usually used for PC back-up purposes.

ting codes to descriptive SGML turned out to be a very complex task. The first step was to acquire documentation on the 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.[6] However, a full and precise description of the macros used for the Debates was 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 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.

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), a pre-processing phase was introduced as preparation to the automatic conversion program.

In practice, the data preparation thus proceeded in a stepwise 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. At this point, the intermediate files were edited to avoid the errors otherwise produced by the automatic conversion routines. 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 comprised 2,500 lines of code.

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, e.g. marking headlines and identifying speakers and languages) cannot be reliably reconstructed without quite intelligent processing. For the ASCII data containing some markup, a set of programs to convert these codes into SGML were developed. The file organization in the completed corpus reflects the logical organization of the sittings according to when they were held. The files are named as follows.

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

where <JJ><MM> are year and month of the debate, <D1> is the first day and <D2> the last day of the meeting. LA is the language code (da=Danish, de=German, en=English, es=Spanish, fr=French, gr=Greek, it=Italian, nl=Dutch, pt=Portuguese). For example, the filename deb940418-22.da.sgm refers to the Danish version of the parliamentary sitting from April 18 to April 22, 1994. The date and language codes correspond to the information given for the printed versions.[7]

2.2 Written Questions

The second parallel corpus 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 (about 1.1 million words per language).

The data was acquired from the Office of Publications of the European Community and consists of material published in 1993. The corpus was supplied with a form of SGML markup defined in the FORMEX specification (Guittet, 1985). This markup has been converted to TEI conformant SGML (Sperberg-McQueen & Burnard, 1994) as described in section 4. The corpus preparation consisted of the following steps:

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

The structure of the converted 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. The naming convention is similar to the debates, `exp. joc$<NNN>.93.<LA>.01.tei`, where NNN is three digits referring to the issue of the Written Questions.

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.

[6] 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".

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

Initially, we created a list of possible newspapers for each language. Having chosen the newspapers, we then contacted the newspapers asking whether 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.[8] However, negotiation of licences and delivery of data took somewhat more time. Once we received the corpus data, SGML-markup proceeded quite smoothly (cf. the MLCC final report for details). A brief summary of the data provided **by** the six newspapers are listed below.

Dutch - Het Financiell Dagblad -1992-1993

The corpus contains articles from the Dutch financial newspaper Het Financiell Dagblad editions of 2nd January 1992 through to 24th December 1993. It contains around 8,5 million words of text.

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.

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.

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 most of the others.

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 comprises some 1.88 million words. The data was obtained from the newspaper by PISA University.

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. Like with the Handelsblatt, there is no overlap to the time periods of the other texts. It comprises some 10 million words.

4 SGML-Markup of the Data

In this section we discuss some of the technical issues which we encountered when SGML-encoding the MLCC corpus. A more detailed description can be found in the MLCC final report (Armstrong, et al., 1995).

4.1 Character set

The corpora, as we received them, were coded in ASCII but used several idiosyncratic codings for non ASCII characters. In order to obtain consistency, we recoded the corpora using the well-defined ISO-Latin (ISO 8859) set of character encodings. All of these character sets have 256 characters, i.e. they use 8 bits per character. They are also all identical to ASCII for the first 128 character codes. For all languages with the exception of Greek we have used ISO-Latin-1. For Greek we used ISO-Latin-7. Where characters appear which are not in these character sets we used SGML entities, a complete list of which can be found in the MLCC final report.

4.2 Markup scheme

The choice of the mark-up scheme is based on the Text Encoding Initiative (cf. Sperberg-McQueen & Burnard, 1994) which is an application of SGML taking into account the needs of researchers interested in literary and linguistic corpus-based studies. SGML is a standard for text markup which - particularly in the modified form of HTML for the web - has become the de-facto standard. The TEI is a proposed standard for the markup of text corpora. It defines a set of SGML document type descriptions (DTDs) and a set of markup guidelines for this purpose.

The 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. Additional markup was provided in cases where it could be easily extracted from the original markup, e.g. quoted sections, rendition information and some abbreviations, names and dates.

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 linguistic tools. A file is in nSGML format if it satisfies the following conditions:

- The document is a valid SGML document according to some supplied DTD.
- The document is coded in ISO-LATIN character sets, with embedded character entities as necessary.
- Reference concrete syntax
- processing 8-bit clean in data and attribute values.
- No capacity/length restrictions.
- No short refs or tag minimisation.
- No SUBDOCs.
- No marked sections.
- All end-tags present (except for empty elements).
- All entity references terminated with « ; »
- No SGML elements are broken across multiple lines.

The MLCC corpora are in nSGML format as defined above. The data has been checked for SGML conformance in according to the DTDs elaborated for the sub-corpora. The majority of data in the MLCC corpus is conformant to the TEI P3 DTD which were elaborated for the sub-corpora. All of data were run through the SGMLS parser successfully. This means that all SGML markup in this corpus has been validated, a fact which is not always true of corpora claiming to be SGML marked up.

The document type definition used in preparation and checking of the subcorpora accounts for all markup specific to this corpus (that was recoverable from the original formatting). The header elements were automatically

[8] We would like to thank the editorial boards of the newspapers for their generosity in donating data to this project and their help in providing information about the data and its markup

generated by program and filled in with a small amount of document specific data.

4.3 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 in the numerous files does require 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. We offer these comments as input to future data preparation projects. The problems 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 project an estimate was made of time necessary for the work foreseen. Practical obstacles as discussed above required some readjustment of plans. All of the data is currently in a form useful for NLP activities.

Investment in data conversion work is a trade-off between time and accuracy. Older typesetting languages are procedural languages and 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. 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, some non-reversible step were deemed necessary. These are documented and intermediate copies of the data are 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.

5 Conclusions

The MLCC project was an effort to provide large high-quality multilingual corpora in a wider range of European languages than provided by existing multilingual corpora, such as the Hansard corpus, the UN and the ILO corpus. All texts were converted to a common TEI-compatible SGML-format and are readily available to both academic and commercial users. We hope that these resources will be useful as a basis for NLP development. The corpus has already proved valuable as a test case for software and resource development in the MULTEXT project, an LRE project for the development of MULtilingual TEXTs and tools (Armstrong, 1996). A subset of the parallel data in English and French is currently being used for an evaluation exercise in alignment techniques in the ARCADE project (Langlais, et al. 1998) and a larger subset including French, Spanish and Italian will serve as the basis for evaluating sense disambiguation techniques for SENSEVAL sponsored by SIGLEX ("http://www.itri.brighton.ac.uk/events/senseval/cfp.txt").

We would be glad to hear about any other activities that are using the MLCC corpus and to receive feedback on errors that you discover.

REFERENCES

Armstrong, S. (1996). MULTEXT: Multilingual Text Tools and Corpora. In: Helmut Feldweg & Erhard Hinrichs (Eds.), Lexicon and Text. (pp. 107-119). Tubingen: Niemeyer, 1996. pp. 107-119.

Armstrong-Warwick, S. (1994). Acquisition and Exploitation of Textual Resources for NLP. In Proceedings of the International Conference on Building and Sharing of Very Large Scale Knowledge Bases '93(pp. 59-68).Tokyo.

Armstrong-Warwick, S.; Thompson, H.S.; McKelvie, and Petitpierre, D. (1994). Data in Your Language: The ECI Multilingual Corpus I. In: Proceedings of the International Workshop on Sharable Natural Language Resources (pp. 97-106). Nara Japan.

Armstrong, S.; Kempen, M; McKelvie, D.; Petitpierre, D.; Rapp, R.; and Thompson, H.S. (1995). Multilingual Corpora for Cooperation. Final report for subcontract to EU project LRE 61-101 "International Co-operation for EAGLES". ISSCO,HCRC.

Guittet, C. (Ed.) (1985). FORMEX - Formalised exchange of electronic publications. Luxembourg: Office for official publications of the European Communities. 'New technologies - project management' department.

Langlais Ph., Simard M., Theron P., Bonhomme P., Souissi E., Isabelle P., Armstrong S., Debili F., and Véronis J. (1998). The ARC-A2 A Cooperative Research Project on Bilingual Text Alignment (in proceedings).

Liberman, M. (1989). Text on Tap: The ACL/DCI. In Proceedings of the 1989 DARPA Speech and Natural Language Workshop, Cape Cod.

McKelvie, D.; Thompson, H.S. (1994). TEI-Conformant Structural Markup of a Trilingual Parallel Corpus in the ECI Multilingual Corpus 1. Proceedings of the 2nd Workshop on Very Large Corpora, Kyoto.

Sperberg-McQueen, C.M.; Burnard, L. (Eds.) (1994). Guidelines for electronic text encoding and exchange TEIP3). Chicago/Oxford: ACH, ACL, ALLC.