MULTEXT

Project ID: LRE62050
Funded under: FP3-LRE

Multilingual Text Tools and Corpora

From 1994-01-01 to 1996-03-01

**Total cost:**
Not available

**EU contribution:**
Not available

**Objective**

The project seeks to contribute to the development of generally usable software tools to manipulate and analyse text corpora and to create multi-lingual text corpora with structural and linguistic markup. It will attempt to establish conventions for the encoding of such corpora, as well as guidelines for text software development, building on and contributing to the preliminary recommendations of the relevant international and European standardization initiatives.

The consortium is committed to make its results, namely corpora, related tools, specifications and accompanying documentation, freely and publicly available.

Six major European companies are involved in the project as industrial partners. They will both contribute to the specification and development of the basic tools and provide a first indication of the exploitability of these tools by using them as a basis for building several high-level NLP applications.

**Approach and Methodology**

At the outset of the project, the consortium will undertake to analyse, test and extend the SGML based recommendations of the Text Encoding Initiative (TEI) on real-size data, and gradually develop encoding conventions specifically suited to multi-lingual corpora and the needs of NLP and MT corpus-based research. To manipulate large quantities of such texts, the partners will develop conventions for tool construction and use them to build a range of highly language-independent, atomic and extensible software tools.

These specifications will be the basis for the development of two major software resources, namely (a) tools for the linguistic annotation of texts (e.g. segmenters, morphological analysers, part of speech disambiguators, aligners, prosody taggers and post-editing tools), and (b) tools for the exploitation of annotated texts (e.g. tools for indexing, search and retrieval, statistics). This software will be implemented under UNIX, while its specific properties should facilitate portability to other systems. Moreover, it will be integrated by means of a common user interface into a text corpus manipulation system expected to provide the basic functionality needed in academic or industrial corpus research. For the overall software design as well as the development of specific components, MULTEXT will capitalise on the preliminary results achieved in the ALEP project.

By using the emerging software tools, the consortium plans to produce a substantial multilingual corpus, including parallel texts and spoken data, in six EC languages (English, French, Spanish, German, Italian and Dutch). The entire corpus will be marked for gross logical and structural features; a subset of the corpus will be marked and hand-validated for sentence and sub-sentence features, part of speech, alignment of parallel texts, and speech prosody. All markup will have to comply with the TEI-based corpus encoding conventions established within the project. The corpus will also serve as a testbed for the project tools and a resource for future tool development and evaluation.

An application programming interface will facilitate the coupling of the progressively refined software and data components with several existing language application systems or prototypes. In particular, the industrial partners plan to develop extraction software for lexical and terminological information to complement and improve their Terminology Management, Information Retrieval or Machine Translation systems. Some effort will also be devoted to a prototypical application for testing
and comparing successive versions of a Machine Translation system.

Exploitation and Future Prospects

Text-oriented methods and software tools have come to be of primary interest to the NLP community. It is therefore expected that the availability of basic multi-lingual tools and data will improve and extend R&D across a wide range of disciplines, including not only the various areas of NLP (language understanding and generation, translation, etc.), but also fields such as speech technology, language learning, lexicography and lexicology, literary and linguistic computing, information retrieval, etc. By feeding the results into several commercial application systems/prototypes, the project is expected to show the potential of state-of-the-art methods in corpus linguistics for improving industrially relevant language systems and services.

By interacting with prominent research organizations and initiatives inside and outside the EC, it is hoped that MULTEXT’s approach will receive the attention of a wide international forum. In a longer term perspective, it can be anticipated that this project will strengthen the methodological and technological foundations for the uniform representation, annotation and exploitation of textual information.

Related information

<table>
<thead>
<tr>
<th>Report Summaries</th>
<th>MULTEXT: multilingual text tools and corpora</th>
</tr>
</thead>
</table>

Coordinator

Université de Provence
29, Avenue Robert Schuman
13621 Aix-en-Provence Cedex 1

See on map

Participants

CAP Devis Systemhaus KSP GmbH
Erich-Herion-Straße 11-13
70736 Fellbach
Germany

See on map

Digital Equipment
Netherlands
See on map

EUROLANG-SITE
France
See on map

ISSCO
Switzerland

See on map

Rank Xerox Research Centre
6 chemin de Maupertuis
38240 Meylan
France

See on map
Siemens Nixdorf Informationssysteme AG  
München  
Germany  
See on map

Siemens Nixdorf-CDS  
Spain  
See on map

UNIVERSITAT AUTONOMA DE BARCELONA  
Spain  
See on map

UNIVERSITEIT UTRECHT  
Heidelbergaan 8  
Netherlands  
See on map

Universitat Central de Barcelona  
Spain  
See on map

University of Edinburgh  
Edinburgh  
United Kingdom  
See on map

Università degli Studi di Pisa  
Pisa  
Italy  
See on map

WESTFAELISCHE WILHELMS - UNIVERSITAET MUENSTER  
Domagkstrasse 5  
48129 MUNSTER  
Germany  
See on map

Last updated on 1994-01-19
Retrieved on 2019-07-26

© European Union, 2019