MULTEXT : Multilingual Text Tools and Corpora

Nancy Ide and Jean Véronis

LABORATOIRE PAROLE ET LANGAGE
CNRS & Université de Provence
29, Avenue Robert Schuman
13621 Aix-en-Provence Cedex 1 (France)

e-mail: ide@fraix11.univ-aix.fr, veronis@fraix11.univ-aix.fr

Abstract. MULTEXT (Multilingual Text Tools and Corpora) is the largest project funded in the Commission of European Communities Linguistic Research and Engineering Program. The project will contribute to the development of generally usable software tools to manipulate and analyse text corpora and to create multilingual text corpora with structural and linguistic markup. It will attempt to establish conventions for the encoding of such corpora, building on and contributing to the preliminary recommendations of the relevant international and European standardization initiatives. MULTEXT will also work towards establishing a set of guidelines for text software development, which will be widely published in order to enable future development by others. All tools and data developed within the project will be made freely and publicly available.

Keywords. multi-lingual corpora, text markup, text software, corpus annotation.

1. Introduction

Text-oriented methods and software tools have come to be of primary interest to the NLP community. However, existing tools for natural language processing (NLP) and machine translation (MT) corpus-based research are typically embedded in large, non-adaptable systems which are fundamentally incompatible. Little effort has been made to develop software standards, and software reusability is virtually non-existent. As a result, there is a serious lack of generally usable tools to manipulate and analyze text corpora that are widely available for research, especially for multi-lingual applications.

At the same time, the availability of data is hampered by a lack of well-established standards for encoding corpora. Although the Text Encoding Initiative (TEI) has provided guidelines for text encoding [Sper94], they are so far largely untested on real-scale data, especially multi-lingual data. Further, the TEI Guidelines offer a broad range of text encoding solutions serving a variety of disciplines and applications, and are not intended to provide specific guidance for the purposes of NLP and MT corpus-based research.

MULTEXT (Multilingual Text Tools and Corpora) is a recently initiated large-scale project funded under the Commission of European Communities Linguistic Research and Engineering Program,. which is intended to address these problems. The project will contribute to the development of generally usable software tools to manipulate and analyse text corpora and to create multilingual text corpora with structural and linguistic markup. It will attempt to establish conventions for the encoding of such corpora, building on and contributing to the preliminary recommendations of the relevant international and European standardization initiatives. MULTEXT will also work towards establishing a set of guidelines for text software development, which will be widely published in order to enable future development by others. The project consortium, consisting of eight academic and research institutions and six major European industrial partners, is committed to make its results, namely corpus, related tools, specifications and accompanying documentation, freely and publicly available.

2. Project Overview

At the outset of the project, the consortium will undertake to analyse, test and extend the SGML-based recommendations of the 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, in collaboration with the recently established Text Software Initiative (TSI), 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 experience and, possibly, 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 to 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.

3. Background and approach

3.1. Software Standard

MULTEXT is strongly committed to "software reusability", to avoid the re-inventing of the wheel and development of largely incompatible and non-extensible software that is characteristic of much language-analytic research in the past three decades. Therefore, the project will establish a *software standard* for the development of its tools. This will enable these tools to be universally used and extended by others.

We outline here the principles (borrowed from [IdeV93a]) underlying the MULTEXT approach to software design, which enable flexibility, extendability, and reusability.

• *Principle 1: Language independence*

The first goal is to extend existing methods to other European languages. So far, these methods have been applied almost exclusively to English. Therefore, the methods will be adapted to produce language-independent tools, by using an *engine-based* approach where all language dependent materials are provided as data. Thus, extension of the tools to cover additional languages will in most cases involve only providing the appropriate tables and rules.

• *Principle 2: Atomicity*

Existing text analytic software often comprises large, integrated systems that are nearly impossible to adapt or extend. MULTEXT will produce a set of small tools (often on the order of a few lines of code, with the absolute minimum of functionality) that researchers can use alone or combine to create larger, more complex programs, thereby implementing a "software Lego" approach. In this way, increasingly complex program bundles can be developed without the overhead of large system design, and with ease of modification since any program can be de-bundled into its constituent programs, each consisting of small, easily understandable piece of code. MULTEXT will bundle its tools in a comprehensive corpus-handling system, as well as demonstrate their use in several high-level applications, thus showing different ways in which the "Legos" can be recombined in specific applications.

• *Principle 3: Operator/stream approach*

MULTEXT will adopt the operator/stream approach to software design, which has had widespread implementation and use and is generally accepted in research and industry. In particular, it has been used increasingly in computational linguistics applications (see, for instance, [Libe92]). The operator/stream approach has served as the basis for the UNIX operating system, which as a result provides a ready-made platform for its implementation.

In the operator/stream approach, data flows in uni-directional "streams" between functions. Each of these functions is an "operator" that transforms the data as it passes by. Since everything is understood in terms of what goes in and what comes out, the emphasis is on what needs to be done rather than how it is done. This enables a focus on overall algorithms rather than implementation details. Component functions are independent, and at no point are compiled together in a single program. This is a key point, since it means that each operator can be implemented in a different language, developed by different people, tested independently, etc. In addition, new functions can be plugged into the stream as needed, and all functions are completely re-usable in other contexts.

• *Principle 4: Unique data type*

Communication between programs will be by means of flat, human readable streams and files, apart from well-defined, encapsulated binary formats for cases such as speech signal, images, or indexes. The only data type is therefore the string. There is some overhead in this approach, since conversion from string to, say, number and back is required for numbers that are to be manipulated arithmetically, but the speed and storage capacities of present-day machines virtually eliminate this concern. More importantly, the use of string data only enables an easy test-modify-test cycle, since the input and output of any step can be examined and manipulated using all-purpose tools freely available on most machines, such as text editors, search software, sorting utilities, etc. Finally, complex data types tie programs to specific languages that implement those types. The use of a unique data type eliminates this dependency.

A feature of this strategy which is of major importance is that any system can accept flat files. Therefore, data is portable between different systems. In addition, it is much easier to port software from system to system, since the software accepts the same kind of input data. For example, a program in C is likely to work on any system with no or very minor modification.

- *Principle 5: Internal standard formats (ISFs)*

To write the compatible set of tools we describe, it is essential that all programs communicate effectively. This demands that internal standard formats (ISFs) for data be developed, to serve as specifications for program development. It is essential that these formats are public, so that any program written anywhere by anyone can use them.

ISFs, like the functions that process them, are very simple and straightforward. Many ISFs will be needed to accomodate different possible "interpretations" of the data, and their development will demand careful consideration of text types, their structures and properties. Therefore, ISF development should build upon the TEI's work on text structures and categories and ensure compatibility with it. Note that because ISFs represent only partially the information in an encoded text (that is, whatever is required for certain operations), they do not replace a TEI/SGML encoding of data, which represents *all* the information in an encoded text and can be used for interchange. Transduction programs to import TEI-conformant texts into one or more internal standard formats, and vice versa, will be essential.

3.2. Tools

All MULTEXT tools will be developed according to the principles outlined above. The project will use only well-known, state-of-the-art methods in tool development, in order to ensure the project's feasibility (e.g., [Chur88], [Cutt92], [Gale91], [Hirst93], [Hirst91]). The project will use these methods to produce a set of tools that is freely available, coherent, extensible, and language independent. The tools will be implemented under UNIX, but will be developed according to principles that will facilitate portability to other systems.

The high-level tools produced by the project fall in two general categories of corpus-handling functions that are basic across applications (these functions apply to mono-lingual texts, multi-lingual parallel texts, and speech):

- *Corpus annotation tools:*
  - segmenter: marks sentences, quotations, words, abbreviations, names, terms, etc.;
  - morphological analyser: provides possible lemmas, morpholgical features, and parts of speech;
  - part of speech disambiguator: disambiguates part of speech where alternatives exist;
  - aligner: provides alignments of passages among parallel texts;
  - prosody tagger: derives automatic modelling of F0 curve and symbolic coding of intonation from the speech signal;
  - post-editing tools: assist in hand validation of automatically annotated corpora.

- *Corpus exploitation tools:*
  - indexing tools: construct indexes for fast access to data;
  - search and retrieval tools: browsing, concordancing, retrieval of collocations, etc., based on a given word, words, pattern, syntactic category, etc.;
  - statistical and quantitative tools: generate lists and statistics--basic statistics for words, collocates (pattern or part of speech) such as frequency, mutual information, etc. Also word lists, lists by syntactic category, etc.

To provide support for these tools, several other general utilities will be required, such as general data manipulation tools, UNIX shell tool, etc. In addition, the tools will be integrated by means of a common user interface into a general-purpose corpus manipulation system suitable for NLP and MT research.

3.3. Markup Standard

One of the goals of MULTEXT is to develop standards for encoding text corpora.

We distinguish four levels of document markup:

- *Level 0. Document-wide markup:*
  - bibliographic description of the document, etc.
  - character sets and entities
  - description of encoding conventions
- *Level 1. Gross structural markup:*
  - structural units of text, such as volume, chapter, etc., down to the level of paragraph
  - footnotes, titles, headings, tables, figures, etc.
- *Level 2. Markup for sub-paragraph structures:*
  - sentences, quotations
  - words
  - abbreviations, names, dates, terms, cited words, etc.
- *Level 3. Markup for linguistic annotation:*
  - morphological information
  - syntactic information--e.g., part of speech
  - alignment of parallel texts
  - prosody

Level 0 provides global information about the text, its content, and its encoding. Level 1 includes universal text elements down to the level of paragraph, which is the smallest unit that can be identified language-independently. Level 2 explictly marks sub-paragraph structures which are usually signalled (sometimes ambiguously) by typography in the text and which are language dependent. Level 3 enriches the text with the results of some linguistic analyses.

The TEI guidelines [Sper94] provide the basis for MULTEXT corpus markup for levels 0 (the TEI header), 1 and 2 as well as many elements of level 3. However, the TEI standard will need careful examination and adaptation [IdeV93b]:

(1) the TEI scheme is intended to be maximally applicable to a variety of encoding purposes and applications. Therefore it in many cases specifies several encoding options for the same phenomena, and provides options and elements without the specific needs of corpus markup in mind.

(2) the TEI scheme is not complete; many areas are yet to be addressed. For example, no TEI encoding scheme

for some aspects of spoken materials, such as prosody (F0 modelling, symbolic coding, etc.), exists.

(3) the TEI scheme is largely untested on corpora, especially multi-lingual corpora. Therefore, use of the TEI scheme for corpus encoding will almost certainly require modification and extension. For instance, TEI mechanisms for alignment will require extension and/or modification to handle multi-lingual text alignment and alignment of different levels of speech representation (signal, orthographic transcription, phonemic transcription, prosody).

(4) the TEI scheme specifically does not aim to provide recommendations for certain content-related elements. For example, while the TEI provides several means to *mark* POS, it is not within the scope of the TEI to provide a standardized set of POS category *names*. Instead, it provides a flexible mechanism that can accomodate any set of actual tag names. Similarly, the TEI does not provide guidelines for names which might, for example, be used as identifiers for texts, text categories, etc.

MULTEXT will use the TEI scheme as the basis for the development of a TEI-conformant *Corpus Encoding Style (CES)* that is optimally suited to NLP research and can therefore serve as a widely accepted TEI-based style for European corpus work.

3.4. Corpus

The goal of MULTEXT is not to duplicate the various large multi-lingual data gathering initiatives by collecting raw data. The intent of the project is to provide a valuable resource that is not provided elsewhere, in the form of a high quality multi-lingual corpus for six European languages, annotated for basic structural features as well as sub-paragraph segmentation, POS, and alignment of parallel texts.

The primary goal of the MULTEXT corpus is to provide an example and testbed for:

(1) multi-lingual tools (especially engine-based tools, alignment software, and multi-lingual extraction tools); and

(2) markup across a large variety of languages (including TEI text markup and the NERC pan-european part-of-speech tagset [Mona92]).

MULTEXT has a secondary but important goal to provide a corpus of value for general linguistic analytic purposes, and will aim to serve this goal to the extent possible without compromising or complicating the primary goal.

The corpus will aim for three parts, each comprising six languages (English, French, German, Italian, Spanish, Dutch):

(1) a *comparable corpus*, consisting of 2M words per language, composed of comparable types of texts from two or three different domains. Ten percent of the corpus for each language will be marked and hand validated for sub-paragraph segmentation and POS.

(2) a *parallel corpus*, composed of fully parallel texts across the six languages and including 2M words per language. Half of the corpus for each language will be marked and hand-validated for sentence alignment. Ten percent of the corpus for each language will be marked and hand-validated for sub-paragraph segmentation and POS.

(3) a small *speech corpus*, consisting of additional markup to be used in conjunction with the EUROM-1 speech database. There is movement towards the integration of NLP and speech (see, for example, ELSNET); MULTEXT will explore the possibilities for such integration by attempting to harmonize tools and methods from both areas. MULTEXT will pay special attention to phenomena at the intersection of the two domains, in particular prosody, whose supra-segmental nature invites research into the complex relationships it holds with morphology and syntax.

To serve its goals, MULTEXT will aim to construct its corpus according to the following principles:

- *Principle 1: Consistency*

The same six languages will be represented in equal amounts in all parts of the corpus. Similarly, equal amounts of the same types of texts will be provided for each language.

- *Principle 2: Variety rather than representativeness*

The MULTEXT corpus is small-scale compared to national efforts aimed at providing balanced, representative corpora in a single language. The project does not therefore aim at representativeness or balance in constructing its corpus. Instead, the MULTEXT corpus will contain a *variety* of texts of different types and from different domains, generally following (where appropriate) known criteria from corpus linguistics.

- *Principle 3: High quality of markup*

In the state of the art, automatic markup of segmentation, POS, and alignment is about 90-96% correct for English (and French in the case of the Hansard). In order to provide a reference corpus for further testing of methodologies and tools, MULTEXT will hand-validate a portion of its corpus to make it virtually error-free.

- *Principle 4: Reuse of available data*

MULTEXT is not committed to the goal of collecting data, but rather to enhancing with structural and linguistic annotation data which may be available from other sources. The project therefore aims to use existing, clean data to the extent possible, in order to avoid the overhead of the acquisition process.

- *Principle 5: Commitment to standards*

MULTEXT will use, build upon, and contribute to standards for text markup, including those of the TEI as well as the EAGLES pan-European POS tagset. Because neither of these schemes have been widely tested, the MULTEXT corpus will provide both a testbed and a basis for their evaluation and modification or extension.

4. Exploitation and Future Prospects

It is 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 applications 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.

References

[IdeV93a] Ide, N., Veronis, J. (1993). What next after the Text Encoding Initiative? The need for text software. *ACH Newsletter,* Winter 1993, 1-12.

[Libe92] Liberman, M., Marcus, M. (1992). *Tutorial on Text Corpora,* Association for Computational Linguistics Annual Conference.

[Mona92] Monachini, M., Ostling, A. (1992). *Towards a Minimal Standard for Morphosyntactic Corpus Annotation,* Report of the Network of European Reference Corpora, Workpackage 8.2.

[Chur88] Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted texts. In *Proceedings of the Second Conference on Applied Natural Language Processing.* Austin, Texas, 136-143.

[Cutt92] Cutting, D., Kupiec, J., Pedersen, J., Sibun, P. (1992). A Practical Part of Speech Tagger, *Proceedings of the Third International Conference on Applied Natural Language Processing,* Trento, 133-140.

[Gale91] Gale, W., Church, K.W. (1991). A Program for Aligning Sentences in Bilingual Corpora, *Proceedings of the ACL Conference,* Berkeley, 177-184.

[Hirst93] Hirst, D., Espesser, R. (1993) Automatic modelling of fundamental frequency. *Travaux de l'Institut de Phonetique d'Aix,* 15, 71-85.

[Hirst91] Hirst, D., Nicolas, P., Espesser, R. (1991) Coding the F0 of a continuous text in French : an Experimental Approach. *12eme Congres International des Sciences Phonetiques,* Aix-en-Provence, 5, 234-237.

[IdeV93b] Ide, N., Véronis, J. (1993). Background and context for the development of a Corpus Encoding Standard, *EAGLES Working Paper,* 30p.

[Sper94] Sperberg-McQueen, C. M., Burnard, L. (1994) *Guidelines for Electronic Text Encoding and Interchange,* Text Encoding Initiative, Chicago and Oxford (in press).

Appendix - Descriptive overview

MULTEXT (Multilingual Text Tools and Corpora)

Coordinator
Dr. Jean Véronis
Laboratoire Parole et Langage
CNRS & Université de Provence
29, Avenue Robert Schuman
F-13621 Aix-en-Provence Cedex 1
tel: +33 42 95 20 73
fax: +33 42 59 50 96
e-mail: veronis@fraix11.univ-aix.fr

Start Date	Jan. 1994
Duration	26 months
Resources	238.5 person-months
Estimated total cost	3.210.000 ECU

Partners	**Country**
CNRS	FR
EUROLANG-SITE	FR
INCYTA	ES
Digital Equipment B.V.	NL
CAP debis Systemhaus KSP	DE
University of Pisa (ILC/CNR)	IT
University of Edinburgh (HCRC/LTG)	UK
ISSCO	CH

Associated Partners	**Country**
Siemens Nixdorf Informationssysteme AG	DE
Universitaet Muenster	DE
Rank Xerox Research Center	FR
Universitat Autonoma de Barcelona	ES
Universitat Central de Barcelona (FBG)	ES
Universiteit Utrecht	NL