NL and Speech in the MULTEXT Project

Jean Véronis, Daniel Hirst, Robert Espesser, Nancy Ide

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

e-mail: `veronis@fraix11.univ-aix.fr`

Abstract. MULTEXT is the largest project funded under the LRE Program, intended to contribute to the development of generally usable software tools to manipulate and analyse multi-lingual text and speech, and to annotate multi-lingual text and speech 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 linguistic software development, which will be widely published in order to enable future development by others. The place of speech in the overall project is intended to explore the possibilities of integrating NL and speech processing by attempting to harmonize tools and methods from both areas. MULTEXT will focus on 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.

1. Project Overview

MULTEXT is the largest project funded under the European Commission's LRE (Linguistic Research and Engineering) Program, intended to contribute to the development of generally usable software tools to manipulate and analyse multi-lingual text and speech, and to annotate multi-lingual text and speech 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 linguistic 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, tools, specifications and accompanying documentation, freely and publicly available. The place of speech in the overall project is intended to explore the possibilities of integrating NL and speech processing by attempting to harmonize tools and methods from both areas. MULTEXT will focus on 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.

At the outset of the project, the consortium will (in cooperation with the European Advisory Group on Language Engineering Standards, EAGLES) 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 NL and Speech corpus-based research. To manipulate large corpora, 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 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.

By using the emerging software tools, the consortium plans to produce a substantial annotated 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; subsets of the corpus will be marked and hand-validated for sentence and sub-sentence features, part of speech, alignment of parallel texts, and 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.

2. Deliverables

2.1. 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 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 linguistic software often comprises large, integrated systems that are nearly impossible to adapt or extend. MULTEXT will produce a set of small tools 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 a 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 NL and speech 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.

• *Principle 4: 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. Whenever possible , these ISFs will be flat, human-readable, not binary. It is essential that these formats are public, so that any program written anywhere by anyone can use them.

2.2. 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], [Hirst91], [Hirst93]). 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:*

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

• *Corpus exploitation tools:*

(1) indexing tools: construct indexes for fast access to data;
(2) search and retrieval tools: browsing, concordancing, retrieval of collocations, etc., based on a given word, words, part-of-speech category and patterns, prosody patterns, etc.;
(3) 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.

The tools will be integrated by means of a common user interface into a general-purpose corpus manipulation system suitable for NL and Speech research.

2.3. Markup Standard

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

We distinguish four levels of document markup:

•*Level 0. Document-wide markup:*

(1) bibliographic description of the document, etc.
(2) character sets and entities
(3) description of encoding conventions
(4) recording conditions, etc.

• *Level 1. Gross structural markup:*

(1) structural units of text, such as volume, chapter, etc., down to the level of paragraph
(2) footnotes, titles, headings, tables, figures, etc.

• *Level 2. Sub-paragraph structures:*

(1) sentences, quotations
(2) words
(3) abbreviations, names, dates, terms, cited words, etc.

• *Level 3. Linguistic annotation:*

(1) morphological information
(2) syntactic information--e.g., part of speech
(3) alignment of parallel texts
(4) 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], since it 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. MULTEXT will use the TEI scheme as the basis for the development of a TEIconformant *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.

2.4. 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 markup and the EAGLES morphosyntactic recommendations [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 *speech corpus,* consisting of additional markup to be used in conjunction with the EUROM-1 speech database. The speech corpus will be annotated with FO modeling, intonation symbolic coding and minimal alignment of speech signal with word boundaries and stressed syllables.

3. Intersection of NL and Speech: Prosody

3.1. Context

MULTEXT will explore the possibilities for integration of NLP and speech 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.

Research in this area is very important for applications such as high quality text-to-speech synthesis. High quality text-to-speech systems are needed for a wide range of applications: access to files by telephone, aid to handicapped persons, talking books, multi-media man-machine communication, etc. However, two main problems hamper their broad diffusion:

- most systems are oriented towards English synthesis, or a few major EU languages;
- most systems suffer from a lack of synthetic voice quality, in particular a low quality of the generated prosody.

Prosody generation covers various aspects such as stress, intonation and rythm. POS disambiguation is very important for stress assignment (compare *an 'escort* and *to es'cort*), and strong input from the POS disambiguator. Intonation and rythm obviously depend on morphosyntactic factors: pauses occur at boundaries between certain types of phrases, intonation targets are placed on certain types of words. Although a complete syntactic analysis of sentences is highly desirable, it is not achievable on running text with the current technologies. However, preliminary experiments show that the recognition of POS (in particular major vs minor categories) and patterns of POS can improve quality in a drastic way.

MULTEXT will contribute to the field by providing prosody tools that will automatically derive a symbolic representation of the intonation from the speech signal. An automatic modeling system is highly desirable for a number of reasons. First, an efficient tool will be extremely useful for collecting data that can be used to improve both speech synthesis and speech recognition. Second, a symbolic coding will enable vastly reducing the amount of data stored (a few symbols instead of the complete acoustic curve). Most importantly, such a tool will be extremely valuable for testing models of intonation and their relationship with morphology and syntax, as well as examining variability in prosodic parameters across individuals and languages.

3.2. Prosody tools

We have developed a number of tools for analysis which will be used as a starting point in MULTEXT. These tools will be adapted to the general MULTEXT software philosophy and integrated into the general toolkit. The tools will perform three tasks:

(1) Automatic modeling of the F0 curve from the speech signal. This tool will implement the method described in [Hirst91] and [Hirst93], using a technique called "asymmetrical modal quadratic regression". The output of this tool will be a sequence of target points (Hz, ms), which consitute a stylisation of the F0 curve. This output is of great interest in itself, since it has been shown that fundamental frequency synthesis by a quadratic spline function interpolating the target points is virtually indiscernible from the original.

(2) Automatic symbolic coding of intonation from the sequence of target points. This tool will take the sequence of target points generated by the previous tool, and produce a symbolic coding of intonation in the INTSINT system ([Hirst94a], [Hirst94b]). The output will be a sequence of INTSINT symbols (such as Higher, Lower, Same, Downstep, Upstep, Top, Bottom, etc.), coupled with their place of occurrence on the temporal scale (in ms).

(3) Alignment of INTSINT coding and phonemic or orthographic transcription. If the alignment of the phonemic or orthographic transcription with the speech signal is known, at least in terms of word boundaries and stressed syllables, a third tool can easily align the symbolic coding and the transcription.

Since MULTEXT provides tools for marking morphology and POS in the orthographic transcription, this Task makes possible a complete chain of alignments for several levels of language, from the speech signal up through the level of POS.

3.3. Prosody markup

MULTEXT will enhance with linguistic annotation material from the EUROM-1 CD-ROMS produced by the ESPRIT-2 project SAM, comprising 40 short passages of 5 thematically connected sentences, each recorded by several native speakers, for all the MULTEXT languages except Spanish. These CD-ROMS are expected to be available by the Autumn (3 CD Roms per language, 24 in all). As part of the SAM-A project, EUROM-1 recordings are also being carried out in Spanish and are expected to be ready by autumn, 1993. The EUROM-1 CD ROMs also contain the phonemic and orthographic transcriptions of the recordings.

MULTEXT will enhance the EUROM-1 corpus with markup for prosody, segmentation, and POS. The prosody markup will consist of two levels: F0 curve modeling and symbolic coding. This markup will be accomplished using the prosody tools, and hand-validated. The orthographic transcriptions will be marked for level 2 and POS, and hand-validated. This requires only minimal effort since the corpus contains only 200 sentences.

EUROM-1 does not provide a phoneme level segmentation of the signal. It is not within the scope of MULTEXT to carry out this task, and we can expect that before MULTEXT is completed in 2 years, this work will be taken up by other groups (cf. the DK_SALA programme developed in Denmark within SAM [AndD82]). We propose to carry out a restricted alignment, consisting of the alignment of word boundaries as well as the beginning of accented vowels between signal and transcription. We anticipate that this segmental alignment will be sufficient for at least a first approximation for prosody. This alignment will also allow the assessment of convergence between our transcription and the TOBI transcription system, which refers to both word boundaries and the position of stressed syllables [Silv92]. The restricted alignment will be carried out for one speaker/language.

The marking of alignment for several levels of language analysis (signal, phonemic transcription, orthographic transcription, F0 modeling, intonation symbolic coding, word boundaries, POS tagging) will provide a robust test of the TEI alignment mechanisms as well as a challenging problem for the corpus retrieval tools.

The current proposal is to validate the entire speech corpus; however, because such validation has never been done before, an evaluation of the effort required will be carried out after 200 sentences, and (if necessary) this proposal will be adjusted.

This task will produce :

(1) FO modeling of the multi-lingual speech corpus;
(2) intonation symbolic coding of the multi-lingual speech corpus;
(3) minimal alignment of speech signal with word boundaries and stressed syllables, for one speaker per language.

In the process of hand-validation, a complete survey of the number and nature of the prosody tools errors will be made, taking into account differences among languages. A report of this behavior will be distributed with the tools.

4. Conclusion

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

[AndD92] Andersen, O., Dalsgaard, P. (1992) *DK_SALA V1.1 User's Guide* Esprit Project 2859 (SAM), Document no. SAM-IES-059.

[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.

[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.

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

[Hirst94a] Hirst, D., Di Cristo, A. (1994) A survey of intonation systems. In Hirst, D., Di Cristo, A. (eds) (1994) *Intonation Systems: a survey of twenty languages.* Cambridge University Press, in press.

[Hirst94b] Hirst, D., Di Cristo, A. (eds) (1994) *Intonation Systems: a survey of twenty languages.* Cambridge University Press, in press.

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

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

[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.

[Silv92] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. (1992) TOBI : a standard for labeling English prosody. Proc. Internal. Conf. Spoken Language Processing Vol. 2, 867-870.

[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	
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