DOCUMENT RESUME

ED 413 746 FL 024 767

AUTHOR	Erjavec, Tomaz; Ide, Nancy; Petkevic, Vladimir; Veronis, Jean
TITLE	MULTEXT-EAST: Multilingual Text Tools and Corpora for Central and Eastern European Languages.
PUB DATE	1995-00-00
NOTE	12p.; In: Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995); see FL 024 759.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Bulgarian; *Computational Linguistics; *Computer Software; Czech; Descriptive Linguistics; *Discourse Analysis; Estonian; Foreign Countries; Hungarian; Language Research; Linguistic Theory; Program Descriptions; Romanian; Slovenian; Structural Analysis (Linguistics); *Uncommonly Taught Languages
IDENTIFIERS	Europe (Central); Europe (East); European Union; *Language Corpora; *MULTEXT; MULTEXT EAST

ABSTRACT

MULTEXT is a European Union project to identify and develop language resources, language-related software, and standards to make the resources maximally usable. MULTEXT-EAST is a spinoff project to develop significant resources for six Central and Eastern European (CEE) languages (Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovenian) and adapt existing tools and standards to them. MULTEXT has developed a corpus encoding standard (CES), and MULTEXT-EAST is applying it to texts in the six languages. This has led to major revision of the CES, particularly to accommodate additional character sets. MULTEXT-EAST is building an annotated multilingual corpus composed of materials comparable to MULTEXT's, including: (1) at least 100,000 words of fiction and newspaper text in each of the CEE languages; (2) parallel translations of the same fictional text; and (3) a small corpus of spoken texts in each language. MULTEXT-EAST has adapted and extended MULTEXT language-dependent materials (lexicons, morphological rules, etc.) for its six languages. Guidelines for linguistic software development are also in progress. A list of participating organizations is appended. (MSE)

MULTEXT-EAST: Multilingual Text Tools and Corpora for Central and Eastern European Languages

Tomaž Erjavec[*], Nancy Ide[**], Vladimír Petkevič[***], Jean Véronis[**]

[*] Laboratory for Language and Speech Technologies
Institute Jožef Stefan
Jamova 39
61111 Ljubljana
Slovenia
Tel.: +386 61 1773 ext. 507
Fax: +386 61 219 385
E-mail: tomaz.erjavec@ijs.si

[**] Laboratoire Parole et Langage
Centre National de la Recherche Scientifique et Université de Provence
29, Av. Robert Schuman
F-13621 Aix-en-Provence Cedex 1
France
Tel.: +33 42 204 356
Fax: +33 42 205 905
E-mail: veronis@cs.vassar.edu

[***] Institute of Theoretical and Computational Linguistics
Faculty of Philosophy, Charles University
Celetná 13
110 00 Praha 1
Czech Republic
E-mail: vladimir.petkevic@ff.cuni.cz

1. Introduction

The language industries rely increasingly heavily on the availability of large-scale language resources, appropriate software tools, and standards to make them maximally reusable. Such resources and tools exist or are under development for most western languages, and efforts to develop standards for corpus encoding and linguistic software development are well underway, in particular in the LRE project MULTEXT, one of the largest EU projects in the domain of language tools and resources (Ide and Véronis, 1994).

However, there have been no comparable efforts for Central and Eastern European (CEE) languages. No large-scale, systematic attempts at corpus collection currently exist (in particular for multilingual, parallel corpora in these languages); tools specifically adapted to corpora in CEE languages are not widely available; and most standardization efforts have not yet taken into account the specific characteristics of CEE languages.

MULTEXT-EAST is a spin-off of the LRE project MULTEXT which is intended to fill these gaps by developing significant resources for six CEE languages (Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovenian) and by adapting existing tools and standards to them. MULTEXT-EAST extends MULTEXT's scope to CEE languages with the following goals:

- test and adaptation of language standards
- development of an annotated multilingual corpus
- development of morpho-lexical resources
- adaptation of the MULTEXT corpus tools.

MULTEXT-EAST began at approximately MULTEXT's mid-point, at a time when MULTEXT's specifications, methods and tools were well-developed enough to extend to additional languages. At the same time, it has been possible to incorporate feedback from application to vastly different language types (especially Slavic and Finno-Ugric) while specifications, methods, and tools are still under development.

Together, MULTEXT and MULTEXT-EAST create a unique network of more than 20 academic research centers and companies, all developing and using common lingware and methodologies for 13 EU and CEE languages. Moreover, MULTEXT-East will also coordinate its efforts in tool adaptation with the TELRI concerted action, esp. with Working Group for Tool Availability. This working group will promote the MULTEXT tools and help in adapting them to the MULTEXT-East languages and different software platforms.

2. Corpus

2.1 Markup

MULTEXT has developed a Corpus Encoding Standard (CES) (Ide and Véronis, 1995b) optimally suited for use in corpus linguistics and language engineering applications, which can serve as a widely accepted set of encoding standards for European corpus work. The standard identifies a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and linguistic information) as well as general architecture (so as to be maximally suited for use in a text database). It also provides encoding conventions for more extensive encoding of linguistic corpora and for linguistic annotation.

The CES is an application of SGML (ISO 8879:1986, Information Processing-Text and Office Systems--Standard Generalized Markup Language). It is based on and in broad agreement with the TEI Guidelines for Electronic Text Encoding and Interchange (Sperberg-McQueen and Burnard, 1994; see also Ide and Véronis, 1995a). The TEI Guidelines were expressly designed to be applicable across a broad range of applications and disciplines; therefore, they treat not only a vast array of textual phenomena, but are also designed with an eye toward the maximum of generality and flexibility. The CES, on the other hand, treats a specific domain and set of applications, and can, therefore, be more restrictive and prescriptive in its specifications. In addition, because the TEI is not complete, there are some areas of importance for corpus encoding that the TEI Guidelines do not cover. Therefore, the first major task in developing the CES has involved evaluating, adapting, selecting from, and extending the TEI Guidelines to meet the specific needs of corpus-based work.

In its present form, the CES provides the following:

- a set of metalanguage level recommendations (particular profile of SGML use, character sets, etc.);
- tagsets and a DTD for documentation of the encoded data;
- tagsets, DTDs, and recommendations for encoding textual data, including written texts across all genres, for the purposes of corpus-based work in language engineering.
- tagsets, DTDs, and recommendations for encoding linguistic annotation, including segmentation, grammatical annotation, and parallel text alignment.

MULTEXT-EAST is applying the CES to texts in six CEE languages, including fiction and newspaper data. The experience of applying the CES to these new languages has led to a major revision and extenstion of the CES, in particular to handle the required additional character sets. In addition, the lack of substantial pre-existing texts in some electronic format in the Eastern European countries and the resulting need to develop many corpora based on printed materials only has made it necessary to consider the kinds of markup that can or should be included and the optimal stages of markup enhancement when corpora are generated in this way.

2.2 Corpus composition

MULTEXT-EAST is building an annotated multilingual corpus, composed of material comparable to MULTEXT, whose primary goal is to provide an example and test-bed for:

- the applicability of MULTEXT's multilingual tools (especially enginebased tools, alignment software, and multilingual extraction tools) to CEE language corpora;
  and
- the applicability to CEE languages of the TEI Guidelines and MULTEXT's TEI-based corpus markup standard, as well as the MULTEXT-EAGLES pan-European lexical specifications and part-of-speech tagset.

The sample corpus is being prepared in TEI-conformant SGML format and annotated for basic structural features as well as sub-paragraph segmentation, part of speech, and alignment of parallel texts.
The sample corpus will be composed of three major parts:

(1) Multilingual Comparable Corpus
For each of the six MULTEXT-EAST languages, the comparable corpus will include two subsets of at least 100 000 words each, consisting of
- fiction, comprising a single novel or excerpts from several novels;
- newspapers.

The data will be comparable across the six languages in terms of the number and size of texts. Selection criteria will be applied to each subset to ensure quality. The entire multilingual comparable corpus is being prepared in CES format, manually or using ad-hoc tools, and will be automatically annotated for tokenization, sentence boundaries, and part of speech annotation using the project tools. For each language, a portion of the corpus will be hand validated.

(2) Multilingual Parallel Corpus
For the six MULTEXT-EAST languages, the parallel corpus will include approximately 100 000 words per language, consisting of translations of Orwell's Nineteen Eighty-Four. The entire multilingual parallel corpus will be prepared in CES conformant format, manually or using ad-hoc tools, and then automatically annotated using the project tools. For each language, half of the corpus will be marked and validated for alignment and sentence boundaries. Alignment will be between the English version and each of the six MULTEXT-EAST languages, thus, constituting six pair-wise alignments. A portion of the corpus will be hand validated.

(3) Multilingual Speech Corpus
MULTEXT-EAST will record a small corpus of spoken texts in each of the six languages, similar to the EUROM-1 speech corpus, comprised of 40 short passages of five thematically connected sentences, each spoken by several native speakers with phonemic and orthographic transcriptions. MULTEXT-EAST will enhance this spoken corpus with markup for prosody, segmentation, and part of speech. The prosody markup will consist of two levels: F0 curve modeling and symbolic coding. This markup will be performed using the tools developed in MULTEXT, and a portion of the corpus will be hand validated. The orthographic transcriptions will be marked for tokenization, sentence boundaries, and part of speech annotation, and they will be hand validated. The project will carry out a restricted alignment, consisting of the alignment of word boundaries as well as the beginning of accented vowels between signal and transcription for one speaker per language.

3. Morpho-lexical resources

An important aspect of tool development in MULTEXT is the engine-based approach, where all language-dependent materials (lexicons, morphological rules, etc.) are provided as data. MULTEXT-EAST, in collaboration with EAGLES, has evaluated, adapted, and extended the specifications (rule format, lexical specifications, corpus tagset, etc.) for the language-dependent material developed in MULTEXT to cover the six MULTEXT-EAST languages (Monachini, 1995). Accomodating the different language families represented among the MULTEXT-EAST languages has demanded substantial assessment and modification of the pre-existing specifications, which were developed for Western European languages only. The work carried out in MULTEXT-EAST has, thus, broadened the base and contributed significantly to defining a universal mechanism for lexical specification.

MULTEXT-EAST is developing the following language-specific resources for use with the various annotation tools:

(1) Segmentation rules. This includes rules describing the form of sentence boundaries, quotations, numbers, punctuation, capitalization, etc.

(2) Special tokens. The language-specific data required by the segmenter includes lists of special tokens (frequent abbreviations and names, titles, patterns for proper names, etc.) with their types.

(3) Morphological rules. The project is providing morphological rules for the MULTEXT-EAST languages, which are needed by the morphological tools. The rules provide exhaustive treatment of inflection and minimal derivation. Each lemma in the lexical lists used by the project (see below) is associated with its part(s) of speech and morphological rules.

(4) Lexical lists. For each of the six MULTEXT-EAST languages, a lexical list containing at least 15 000 lemmas is being developed for use with the morphological analyser. Each entry includes the following information: inflected-form / part of speech / morphological information / lemma. A mapping from the morpho-syntactic information contained in the lexicon to a set of corpus tags (used by the part of speech disambiguator) is also provided according to the MULTEXT tagging model (Véronis and Khouri, 1995).

4. Tools

4.1 Standardization

There is a serious lack of generally usable tools to manipulate and analyze the text and speech corpora and collections that are now becoming widely available. The linguistic software that exists at present only begins to cover growing needs. Industrial software is often expensive or unavailable, and is usually hard to adapt or extend. On the other hand, the substantial body of natural language processing academic software is often experimental, hard to get, hard to install, under-documented, and sometimes unreliable. In both cases, tools are typically embedded in large, non-adaptable systems which are fundamentally incompatible. Worse, there is enormous duplication of effort: it is not at all uncommon for researchers to develop tailor-made systems that replicate much of the functionality of other systems and in turn create programs that cannot be re-used by others and so on in an endless software waste cycle. Although efforts to develop standards for data representation are underway, little effort has been made to develop standards for linguistic software, and software reusability is virtually non-existent.

MULTEXT has joined efforts with the EAGLES sub-group on Tools to address this need by working towards the establishment of Guidelines for Linguistic Software Development (LSD) (Véronis and Ide, 1995). These guidelines specify a general lingware development environment, including recommended standards for all aspects of software development, data representation, linguistic annotation, etc. The establishment of such a set of guidelines enables the interchange of tools and data among researchers and sites, compatibility among tools with potentially diverse functionality, and in general contributes to the creation of reliable, high quality tools.

Standards exist or are being developed in many areas relevant to linguistic software development, including

- character sets
- document encoding
- language and country codes
- application program interfaces
- programming languages
- internationalization and localization of programs
- etc.

Each of these standards covers a small piece of what would serve as a general lingware development environment, but none has been developed with an eye toward the overall coherence of such an environment. The goal of the MULTEXT/EAGLES LSD Guidelines is to bring together existing or emerging de jure or de facto standards sufficient to address the scope of an entire Linguistic Software Development system.

MULTEXT tools are intended to demonstrate many of the basic principles of software development that will be recommended in this environment, including especially atomicity and language-independence. MULTEXT-EAST provides a significant test-bed for the MULTEXT tools, in particular because these principles are aimed towards enabling easy modification and extension to new (and possibly very different) languages.

4.2 Adaptation of Multext tools

MULTEXT is developing a set of corpus manipulation tools that is freely available, coherent, extensible, and language-independent, including:

Morphosyntactic tagging:
- segmenter: marks sentences, quotations, words, abbreviations, names, etc.;
- lexical lookup and morphological analyser: provides lemmas, morphological features, and parts of speech;
- part-of-speech disambiguator: disambiguates parts of speech where alternatives exist;

Parallel text alignement:
- aligner: provides alignments of sentences among parallel texts;

Prosody tagging:
- signal editor and signal analysis utilities (MES)
- prosody tagger (MOMEL): derives automatic modelling of F0 curve and symbolic coding of intonation from the speech signal;

Corpus manipulation tools:
- SGML query language (SgmlQL);
- format conversion utilities;
- multilingual string manipulation library;
- post-editing tools: assist in hand validation of automatically annotated corpora.

The tools are implemented under UNIX. All MULTEXT tools are designed using an engine-based approach where all language-dependent materials are provided as data. Therefore, extension of the tools to cover CEE languages in MULTEXT-EAST primarily involves providing the appropriate tables and rules for these languages. However, some adaptation of the tools is expected, given the potential for new problems which may be posed by these vastly different language types (i.e., languages with heavy inflection, free word order, etc.).

5. Conclusion

MULTEXT-EAST is extending the MULTEXT effort to six CEE languages by adapting MULTEXT's tools, developing linguistic resources for these six languages, and providing a multilingual corpus comparable to the one developed for EU languages within MULTEXT. This will validate and enhance MULTEXT's tools and its software and markup standards. Most importantly, it will enable not only early use of developing standards in CEE countries but also the possibility for feedback as a result of adaptation to a vastly different set of languages.

As in MULTEXT, all of the work within MULTEXT-EAST will be performed in conjunction with EAGLES and the TEI, and thus provide an extension and validation of the work of these initiatives on standardization to a new range of languages. Similarly, like MULTEXT, MULTEXT-EAST will distribute its results, tools, and corpora and linguistic resources for six CEE languages free or at cost by ftp and CD-ROM.

References

Ide, N., and J. Véronis. 1994. "MULTEXT (Multilingual Tools and Corpora)". Proceedings of the 14th International Conference on Computational Linguistics, COLING'94, Kyoto, Japan 1994, 90-96.

Ide, N. and J. Véronis (eds.). 1995a. The Text Encoding Initiative: background and context. Dordrecht: Kluwer Academic Publishers.

Ide, N. and J. Véronis. 1995b. Corpus Encoding Standard. Document MUL/EAG CES1. <URL:http://www.lpl.univ-aix.fr/projects/multext/CES/CES1.html>

Monachini, M. (ed.). 1995. Common Specifications and Notation for Lexicon Encoding of Eastern Languages.

Deliverable 1.1. Multext-East Project COP-106. ftp:////www.lpl.univ-aix.fr/pub/multext/docs/ME1.1.tex

Sperberg-McQueen, C.M. and L. Burnard. 1994. Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford: Text Encoding Initiative.

Véronis, J. and N. Ide. 1995. Guidelines for Linguistic Software Development. Document MUL/EAG LSD2. <URL:http://www.lpl.univ-aix.fr/projects/multext/LSD/LSD2.html>

Véronis, J. and Khouri. 1995. Etiquetage grammatical: mod+le.
<URL:http://www.lpl.univ-aix.fr/projects/multext/LEX/LEX2.html>

Appendix: Project's fact sheet

MULTEXT-EAST Participants

PART	PARTICIPANT'S FULL NAME	CC	R
AIX	Laboratoire Parole et Langage Centre National de la Recherche Scientifique	FR	C
PISA	Istituto di Linguistica Computazionale Consiglio Nazionale delle Ricerche	IT	A
SOFIA	Department of Mathematical Linguistics Institute of Mathematics Bulgarian Academy of Sciences Sofia (Bulgaria)	BU	P
PRAG	Institute of Theoretical and Computational Linguistics Charles University Prague (Czech Republic)	CZ	P
BYLL	BYLL Software, Ltd. Prague (Czech Republic)	CZ	S
TARTU	Laboratory of the Estonian Language Tartu University Tartu (Estonia)	EE	P
BUDA	Linguistic Research Institute Hungarian Academy of Sciences Budapest (Hungary)	HU	P
MORPH	MorphoLogics Budapest (Hungary)	HU	S
BUCHA	Research Institute for Informatics Bucharest (Romania)	RO	P
ICI	ICI Bucharest (Romania)	RO	S
LJUBL	Laboratory for Language and Speech Technologies Institute "Jožef Stefan" Ljubljana (Slovenia)	SI	P
AMEB	AMEBIS Ljubljana (Slovenia)	SI	S

Abbreviations:

PART:	Participant's short name
CC:	Country Code
R:	Role (C- Coordinator, P- Full partner, A- Associate partner, S- Subcontractor)

Effort: 345 person-months
Duration: 24 months
Start state: 1 May 1995

Contact point:
Dr. Jean Véronis (coordinator)
Laboratoire Parole et Langage
CNRS & Universite de Provence
29, Av. Robert Schuman
13621 Aix-en-Provence Cedex 1 (France)
Tel.: +33 42 95 36 34
Fax : +33 42 59 50 96
E-mail : veronis@univ-aix.fr

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: TELRI - Proceedings of the First European Seminar:"Language Resources for Language Technology", Tihany, Hungary, Sept. 15 and 16, 1995

Author(s): Heike Rettig (Ed.)

Corporate Source:

Publication Date: 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☒

Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) ***and*** paper copy.

The sample sticker shown below will be affixed to all **Level 1** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all **Level 2** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

□

Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at **Level 1**.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature:

Printed Name/Position/Title: Norbert Volz, M.A.
TELRI Project Manager

Organization/Address:
Institut für deutsche Sprache
R 5, 6-13 - 68161 Mannheim
Postfach 101621 - 68016 Mannheim

Telephone: +49 621 1581-437

FAX: +49 621 1581-415

E-Mail Address: volz(at)ids-mannheim.de

Date: 28/11/97

(over)