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

East meets West: Producing Multilingual Resources in a European Context

Tomaž Erjavec,[1] Ann Lawson,[2], Laurent Romary[3]

(l) Dept. for Intelligent Systems, Institute Jožef Stefan,
Ljubljana

(2) Abteilung LEXIK, Institut für deutsche Sprache
Mannheim

(3) Equipe Langue & Dialogue, Loria - CNRS
Nancy

Abstract

The EU concerted action TELRI has released a two-volume CD-ROM, which contains multilingual language resources, namely corpora, lexica, and tools for language engineering. This CD-ROM provides harmonised resources for unprecedented numbers and kinds of languages, mainly from non-EU countries, for which such resources still tend to be scarce. The first volume of the CD-ROM includes the aligned text of Plato's Republic in twenty one languages plus other tools and resources, while the second volume contains extended results of the EU MULTEXT-East project, including the aligned and tagged novel '1984' by George Orwell and accompanying lexica in seven languages. The paper presents the CD-ROM, the methods employed in its creation and its prospective uses.

1. Overview

The EU Concerted Action TELRI (Trans-European Language Resources Infrastructure)[1] has released a CD-ROM containing multilingual language resources, namely corpora, lexica and tools for language engineering. This product represents the concrete results of the joint research aspect of the project, which brought together partners across Europe to work together on a close and practical level. The CD-ROM provides standardised resources for a large number of languages, mainly from non-EU countries, for which such resources still tend to be scarce. The making of these resources served to foster the use of existing conventions and recommendations such as TEI (Sperberg-McQueen & Burnard, 1994), EAGLES and MULTEXT (Ide & Veronis, 1994) in Central and Eastern European countries.

The CD-ROM consists of two volumes. The contents of the first volume arose entirely within the TELRI action, and were produced in a process of dissemination of expertise through the work on a common corpus. The second volume contains the results of the EU MULTEXT-East (Multilingual Text Tools and Corpora for Central and Eastern European Languages) project (Erjavec et al., 1996),[2] which collected and applied standards to a large variety of resource types: corpora, lexica and tools. This project finished in 1997 and its results have been enhanced and prepared for CD distribution in the scope of TELRI.

The two volumes are in broad agreement as to the kind of resources they offer, and the type of encoding they use, while they reflect the different organisational structures from which they grew. TELRI was an EU Concerted Action acting on a broad front, with no funds devoted to labour. Most members also had no a priori experience of SGML/TEI and had no dedicated tools available to annotate their texts. So, for example, the TELRI corpus was encoded directly in generic TEI and TEI Lite, rather than trying to emulate more specific schemes, such as the PAROLE specification (Ridings, 1996).

MULTEXT-East, on the other hand, was a Joint Project able to dedicate funds for research and so could be more focused on the kinds of resources it produced. Here the partners shared platforms and tools, could adopt a more unified encoding practice and were involved in the definition of conventions which would be more applicable in a linguistic engineering context. So, for example, the MULTEXT-East corpus was prepared using MULTEXT tools and is encoded in accordance with the Corpus Encoding Specification, CES (Ide et al., 1996; Ide 1998), both developed for such a context. MULTEXT-East also had a clearly defined set of languages that it aimed to produce resources for, namely Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene, with English serving as the 'hub' of the parallel corpus and as the meta-language of the project.

The production of the two volumes was enhanced by the two teams working in cooperation, as they had to solve similar difficulties (scanning, copyright, etc.) and could share know-how through workshops and exchange various software tools. The concrete result of this cooperation is a double CD-ROM, "East meets West: A Compendium of Multilingual Language Resources". In the next sections we give an overview of its contents, that is, the corpora, lexica, and tools contained on the CD-ROM. The paper finishes with a discussion of the CD-ROM's prospective uses.

2. The Corpora

Each volume of the CD-ROM contains an annotated multilingual corpus. The two corpora differ in composition

[1] For more information on the TELRI action see URL *http://www.ids-mannheim.de/telri/.*

[2] For more information on MULTEXT-East see also URL *http://nl.ijs.si/ME/.*

and encoding, reflecting the different practices adopted in their creation.

2.1. The Plato Corpus

The first volume contains a parallel corpus, comprising Plato's "Republic" in twenty one languages. This corpus grew out of the work within the TELRI Working Groups WG7 "Joint Research", WG5 "TELRI Service Pool", and WG4 "Lingware Availability". The "Joint Research" WG was formed with the aim of encouraging collaboration between as many of the TELRI partners as possible. For this reason, it was decided to focus on building a sample parallel corpus of translations of one text. Such corpora are able to furnish researchers with considerable information about language patterning in general, for instance on translational equivalents, collocations and phraseological units. At least initially, the size of this corpus was of less import than the coverage across the languages represented in TELRI, and the active involvement of as many partners as possible.

Plato's "Republic" was chosen as the main text for developing the parallel corpus. As explained above, a single text was chosen for reasons of practicability. A well-known classical text was chosen on the grounds that translations from the original into most, if not all, of the TELRI project languages would be likely to exist. In addition, since none of the TELRI project members is a native speaker of Ancient Greek, no language would be privileged by having the original version in their own language. All the texts produced and examined in the parallel corpus are "target texts", since only the Greek is the "source". Furthermore, the age of the text and hence also many of the translations would, it was hoped, eliminate many troublesome copyright issues.

Each TELRI member sought out suitable translations of the "Republic" in their own language. Translations could be found in almost all the project languages, with the exception of Estonian and Albanian. In several languages, more than one translation was available, such as for English and Czech. Partners tried to find versions of the texts already in electronic form, whenever possible. These could be downloaded from the Internet, bought on CD-ROM or acquired direct from the publishers. If no electronic version was available, partners scanned the texts using OCR scanners. In exceptional cases, when the quality of the type was too poor to be recognised by the scanner, the text was typed in by hand. TELRI was able to offer financial assistance for such technical matters.

The conditions varied greatly from country to country, as was expected. The extremes were on the one hand the English texts, with an abundance of websites and downloadable texts (although also with their own problems) and on the other the situation in, for instance, Latvia. In Latvian, no full translation could be found, so the edited partial translation was encoded and used. In some cases, not only was there no electronic version available, but indeed the translation was out of print and the only copies to be found in libraries or in private ownership. Since scanning requires a flat page and libraries object to their books being torn apart, in these cases the texts were painstakingly keyed in.

Another important issue was that of copyright. Copyright holders were approached for permission to use their texts in academic research work. A letter was prepared by the Group for partners to use when writing to copyright holders to ask permission to use their text in project work. When texts were not or no longer in copyright, because of their age for instance, no permission was sought but the publisher or source of the text was always acknowledged. The situation differed in each country and various individual solutions were worked out.

The choice of text made for some interesting problems. Recognised as a fundamental philosophical text, "The Republic" has been published in different forms over many centuries. Some of the editions published are treated as "standards". However, for the purposes of the project, they proved to be far from standard. Most versions contained markings or annotations of various kinds, but there was little consistency. Initially, we hoped to be able to use these markings to assist the alignment process and encouraged sites scanning or typing in the texts to retain them. It was then decided to use a particular set of markings, and they were manually inserted into those texts which did not already have them.

We then set about encoding the data. While all partners retained a plain-text version of the "Republic", which is also included on the CD-ROM, it was also encoded according to a widely-accepted convention, namely the TEI P3 guidelines (Sperberg-McQueen & Burnard, 1994). These are based upon SGML, a meta-language for encoding electronic texts with information about the structural layout and content of the document, which is of use when analysing the structure of the document, automatically aligning and manipulating it. Most project partners had little or no experience of this particular encoding method at the start of the project, and this proved to be one of the most important results of this part of the TELRI project.

The texts were encoded according to the TEI guidelines using the following conventions: two levels of embedded **<div>** elements have been used to mark-up the structure of the text in books (these were originally scrolls in the Greek text) and subsections (main divisions in the original text). For some of the versions, the second level had to be inferred in comparison with another translation because they had been dropped by the translator. Paragraphs were either marked as **<p>** for "paragraph" or <**lg**> for "line group", depending on whether these were running text or poems. Dialogues have been considered as running text in this corpus; lower level structural mark-up was limited to **<q>** elements to account for the quotations which are quite numerous in the Republic and **<seg>** elements for what could roughly be seen as an orthographic sentence. We dealt with the classical problem of overlapping structures between quotes and sentences by a) allowing the two to be embedded in one another in both directions (hence the use of **<seg>** instead of **<s>,** which does not bear this property in TEI P3), and b) by splitting sentences, when necessary, at points where a quote would start or end in the middle of them. In the perspective of aligning the various translations with one another, such an operation presents the advantage of yielding a more fine grained representation; we used several other elements to cope with different phenomena observed in the various versions of the text, in particular <**milestone**> to indicate the folio marks derived from an old translation and classically used to refer to the text by philologists.

On the basis of this encoding, the corpus has been automatically aligned up to the level of sentences using the hierarchical aligner devised by Bonhomme et al. (1995) and compiled into a series of HTML files of intertwined texts by pairs of languages. Plain ASCII and HTML versions of each text have also been included in the CD-ROM.

All the texts were then uploaded to the Mannheim TELRI ftp site, along with a brief information file detailing the source, status of encoding and person responsible for the version. In this way, all TELRI partners could download the texts for their own use and, if necessary, further encode or alter the text in collaboration with the partner originally responsible. There was thus an interchange between partners, while the integrity of each version was retained.

The Working Group used the texts to test corpus alignment software. The aligning of texts enables the user to compare the translations of particular words or phrases. This is of use for investigating language, training and improving computer-aided translation tools and also for CALL (Computer Assisted Language Learning). For alignment, some segmentation of the texts had to be undertaken. This commonly involved inserting codes to specify where sentence or paragraph boundaries occur. These could then be recognised by the alignment program, along with the markers described above. Many partners have used the parallel texts to examine specific linguistic features of a chosen language pair. This can reflect on the peculiarities of the languages involved and on the nature of translation generally. Much of this research work can also be found on the CD-ROM, with general information about the project.

2.2. The MULTEXT-East Corpus

The second volume of the CD-ROM contains the corpus of the MULTEXT-East project, with further additions due to TELRI. As the corpus is detailed in Erjavec & Ide, (1998), we here only briefly present its composition and encoding.

The corpus is composed of three parts, namely of a multilingual parallel corpus, similar to the "Republic" one, a multilingual comparable corpus and a small multilingual speech corpus. The complete corpus has been marked up with header and structural information (<div>, <p>, etc.) and is encoded in the Corpus Encoding Specification, CES (Ide et al., 1996), a TEI-like encoding scheme.

The parallel corpus contains the novel '1984' by George Orwell in the original, and translations into the six MULTEXT-East languages. In the scope of TELRI, the corpus was extended by four new translations, namely the Lithuanian, Latvian, Serbian, and Russian ones.[3] The parallel corpus has been marked and validated for sentence boundaries (<s>) and alignment. Alignment is between the each of the ten languages and the English version, thus giving ten pair-wise alignments. The alignments themselves are not included in the primary data, but are expressed in separate documents, which contain only ID references to the sentences aligned. The alignments are thus 'flat', i.e. they do not attempt to model the hierarchical structure of the documents. Furthermore, the MULTEXT-East seven-language parallel corpus has been tokenised, and each word token assigned its part-of-speech, or, more accurately, its morphosyntactic description. To arrive at such an annotated corpus involved first developing a harmonised set of lexical specification, as discussed below. It should be noted that this corpus represents the first such effort for most of the languages involved.

The MULTEXT-East corpus further contains a million word multilingual comparable corpus in the six languages of the project. The comparable corpus consist of two parts, the first being 'fiction', comprising either a single novel or excerpts from several novels, and the second the 'news' part, comprising articles from daily newspapers.

Finally, a small parallel speech corpus is also included: it comprises six translations of forty English passages from EUROM/SAM project, and has been recorded and digitised for four of the languages, namely Estonian, Hungarian, Romanian, and Slovene.

3. Lexical Resources

The second volume of the CD-ROM includes substantial harmonised lexical resources for the six languages of the MULTEXT-East project. Again, these resources are detailed in Tufis et al. (1998), so we here give only a brief overview. The MULTEXT-East lexical resources comprise morphosyntactic descriptions, lexica, and, as mentioned above, the Orwell corpus annotated with this information.

The specification for morphosyntactic descriptions follow the Eagles (Monachini & Calzolari, 1995) and MULTEXT (Bel, Calzolari & Monachini (eds.), 1995) specifications and define 14 grammatical categories (PoS), each of them with a number of specific attributes and attribute-values. The tables thus provide a word-level morphosyntactic 'grammar' for the six languages. The morphosyntactic descriptions themselves are written in a compact string notation, with the first character giving the PoS and defining the meaning of the remaining characters. With these characters, the position in the string gives the attribute, and a one letter code its value. Furthermore, the special character ' -' is used when a certain attribute is not applicable, either for the language or for the particular combination of features. So, for example, the string `'Vmipld--y'` denotes PoS:Verb, Type:main, VForm:indicative, Tense:present, Person:first, Number:dual, Gender:not applicable, Voice:not applicable, Negative:yes.

In the lexica of the project, each entry consists of three fields: the word-form, its lemma and its morphosyntactic description. The lexica provide at least 15.000 lemmas for each language, and cover the corpus of the MULTEXT-East project. Except for Estonian and Hungarian, where this is not possible due to the nature of the languages, all the possible word-forms of the lemmas, i.e. full inflectional

[3] Unfortunately, copyright restrictions have prevented us from including the Latvian translation on the CD-ROM.

paradigms, are included in the lexica.

4. Corpus Tools

Both volumes of the CD-ROM also contain a variety of software tools. These are either in the public domain, or have been produced by the project partners.

Of particular importance for both TELRI and the MULTEXT-East projects was the provision of the proper software for editing, testing and aligning the SGML documents produced at the different sites. This has been released in the context of two main packages, namely:

- an SGML editing environment, which is based on Emacs and comprises Lennart Staflin's PSGML mode and James Clark's nsgmls parser. We made the choice to base all our SGML work upon freely available software, such as the above, which could be widely distributed, rather than on more elaborate, but proprietary software like Softquad's Author/Editor which, despite its intrinsic qualities, would have been more difficult to implement. In particular, the flexibility of the Emacs editing environment has allowed us to deliver it with the various DTDs that have been adopted within the two projects, namely the full TEI, TEI Lite and the CES DTDs.
- the XCorpus environment developed at Nancy[4] and which comprises several SGML aware tools for testing, enhancing (e.g. through sentence segmentation) and aligning multilingual documents. In particular, the alignment program is based upon a hierarchical mechanism which iteratively uses the SGML structure in (possibly embedded) <**div**>, <**p**> and <**seg**> (or <**s**>) elements, to compute the final correspondences at the sentence level.

Both these environments, together with some more experimental tools (concordancers, lexical statistics, dictionary look-up etc.) are provided for various platforms such as Solaris, Windows or Macintosh.

5. Distribution

The CD-ROM has been distributed amongst the TELRI and MULTEXT-East partners and related sites throughout Europe and further afield. It is subject to a User Agreement specifying academic use only. The individual text or resource providers can be contacted if an extension of this Agreement is desired. In addition, the CD-ROM has been made available at cost price to all interested parties, and to this end it has been widely advertised within the language engineering community. For more information or to request a copy of the CD-ROM, please see *http://www.ids-mannheim.de/telri/cdrom.html.*

6. Perspectives

Although the nature of the TELRI and MULTEXT-East resources are clearly specific, the work has more wide-ranging consequences. The CD-ROM provides standardised resources for over 20 languages, and tools to produce further compatible resources, or exploit the ones already provided. Its possible applications are thus numerous. It could provide data for lexical studies, in particular those of translation equivalents; for teaching language or translating; or serve as learning data for taggers, aligners, tokenisers, and similar trainable programs. Aligned language pairings are provided which would otherwise be virtually impossible to find.

The CD-ROM uses HTML to structure its contents and provide documentation on the resources. These HTML pages include external HTTP links, pointing to useful environments and documentation. The CD-ROM can thus also serve as a 'primer' for language engineering applications, being especially useful for sites with poor Internet connections.

Arguably the most important result, though less tangible, lies in the sharing of knowledge, expertise and experience through the work and the development of skills for the future. Institutes, indeed countries, with little or no experience in the Language Engineering field have gained expertise in corpus selection, collection, encoding and manipulation. This expertise is now being used to produce new corpora and to encode existing corpus material. The impetus of the production of the CD-ROM is set to continue in present and future work, not least in the framework of the TELRI-II project, beginning Summer 1998.

References

Bel,N., Calzolari,N. and Monachini, M. eds. (1995). Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets, MULTEXT Deliverable D1.6.1B, Pisa.

Bonhomme, P. & L. Romary. (1995). Projet de Concordances Parallèles Lingua : gestion de textes multilingues pour 1'apprentissage des langues, Actes Génie Linguistique, Montpellier.

Erjavec,T, Ide,N.(1998). The MULTEXT-East Corpus. This volume.

Erjavec,T., Ide,N., Petkevič,V. & Véronis,J. (1996). Multext-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages. Proceedings of the First European TELRI Seminar: Language Resources for Language Technology, 87-98.

Ide,N. (1998). Corpus Encoding Standard: SGML guidelines for Encoding Linguistic Corpora. This volume.

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

Ide,N., Priest-Dorman.G. & Véronis,J. (1996). Corpus Encoding Standard. URL: *http://www.cs.vassar.edu/CES/.*

[4] For more information on the XCorpus environment see URL *http://www.loria.fr/projets/XCorpus/.*

Monachini,M. & Calzolari,N. (1995). Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and in Corpora and Application to European Languages. EAGLES document EAG-LSG-T4.6/CSG-T3.2, Pisa.

Ridings,D. (1996). Text representation in PAROLE. Parole MLAP 63-386, Work package 4.1.3

Sperberg-McQueen,C.M. & Burnard,L. (eds.) (1994). Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford.

Tufiş,D., Erjavec,T. & Ide,N.(1998). Standardised Specifications, Development and Assessment of Large Morpho-Lexical Resources for Six Central and Eastern European Languages. This volume.