Building parallel corpora for eContent professionals

M. Gavrilidou, P. Labropoulou, E. Desipri, V. Giouli, V. Antonopoulos, S. Piperidis
Institute for Language and Speech Processing
Epidavrou & Artemidos 6
151 25 Maroussi, Greece.
{maria, penny, elina, voula, vantonop, spip} @ilsp.gr

Abstract

This paper reports on completed work carried out in the framework of the INTERA project, and specifically, on the production of multilingual resources (LRs) for eContent purposes. The paper presents the methodology adopted for the development of the corpus (acquisition and processing of the textual data), discusses the divergence of the initial assumptions from the actual situation met during this procedure, and concludes with a summarization of the problems attested which undermine the viability of multilingual parallel corpora construction.

1 Introduction

INTERA (Integrated European language data Repository Area, Contract 22076Y2C2DMAL2) is an EU-funded project within the eContent framework, aiming at

- building an integrated European Language Resources (LRs) area by connecting existing data centers at regional, national and international level, and
- at proposing "ways and techniques for LRs packaging to make it a profitable and attractive task to eContent professionals"; as an application of this task, the production of multilingual resources, namely parallel corpora and multilingual terminologies extracted from these, is undertaken (INTERA Technical Annex).

This paper focuses on the second aim of the project, presenting the work carried out in the area of parallel corpus production, identifying the steps followed in this process, in order to point out the problematic areas involved in the task and suggest ways of encompassing them.

2 Methodology and specifications

The process usually followed in the LRs production involves the following tasks: (a) identification of user needs and requirements, (b) specifications for the selection, construction and packaging of the LRs, (c) identification of potential sources, (d) construction of the LRs per se, (e) promotion and distribution of the LRs.

Given that INTERA is an eContent project, the target user group defined by the Technical Annex of the project was *eContent professionals and users*; furthermore, it was decided that the LRs to be produced (which would be of interest to this group) would be *parallel corpora* and *multilingual terminological lists*. Finally, the most important objective of the LRs production was the definition of a business model which would be attractive to the abovementioned target group.

The following sections discuss the actual steps taken for the implementation of these requirements.

The target group of eContent players addressed by the project has been further defined as consisting of professionals involved with the:

- production of digital content (authors or publishers)
- Globalization, Internationalization, Localization and Translation (GILT) processes, and
- development of Human Language Technology (HLT) software, ranging from multilingual information retrieval and extraction tools, to content management and Computer-Assisted Translation or Machine Translation solutions.

The next step concerned the identification of user needs and requirements on the basis of the professionals' working habits and processes. This was achieved by exploiting the results of a number of previous initiatives to roadmap the state-of-the-art in multilingual LRs, in combination with new initiatives undertaken in the framework of the project and targeted to the eContent world.

The surveys conducted in the framework of the ENABLER project (Maegaard et al. 2003, Gavrilidou & Desipri 2003) provided insights as to the existence and availability of different types of LRs, language demand, domains of interest, standards, etc. Although ENABLER focused on the LRs developer's point of view, a number of valuable results were elicited. Other surveys, such as those conducted by ELRA and its distribution agency ELDA aiming at determining the needs of users with respect to available and potentially available LRs (http://www.elra.info/), or surveys available over the Internet through the sites of international organizations such as LISA and IDC or consultancy firms (http://www.globalsight.com, LISA 2001, LISA/AIIM 2001, LISA/OSCAR 2003) shed a light as to the availability of resources and relevant tools.

The information elicited from these surveys was coupled by a study of the activities of the eContent professionals as regards LRs, conducted in the framework of INTERA (Gavrilidou et al, 2004) through the circulation of a questionnaire distributed to potential users, as well as through personal contacts with a number of actors in the relevant fields. The main areas of the study concerned the types of LRs the eContent professionals are interested in, domains and languages of interest, and, most important, policies concerning the way they acquire, use and exploit LRs and tools.

The study of the target group yielded the following specifications:

- *domains*: it is obvious that eContent users are more interested in specialized domains than in general language resources; moreover, the survey results showed health/medicine, tourism, education, law, automotive industry and IT/telecommunications, as being the prevailing ones. In the framework of the INTERA project, however, we decided to focus on the prevailing domains as long as they promote multilingual and multicultural content. The selected domains are: *health*, *tourism*, *education and law*, which correspond to the predominant digital activities, namely, eTourism, eHealth, eLearning, eGovernment and eCommerce.
- *languages*: the focus of eContent and the needs of the users pointed towards the less widely spoken languages, including Balkan and Central and Eastern European languages (i.e the languages of the new EU countries).

  The project aims at the construction of a multilingual parallel corpus of 12 million words in total. The ideal scenario for the intended application of term extraction would be that of having a corpus with a source or pivot language and translations of the same texts in a number of target languages; however, given that the project aims at proposing realistic solutions to be adopted in the future by prospective LRs creators, real-life drawbacks should be taken into account; therefore, the limitations in the availability of existing resources (see section 3.1) dictated the decision to collect resources for four *pairs of languages*: Greek-English, Bulgarian-English, Slovene-English and Serbian-English.

The specifications for the processing of the corpus have been based on the requirements of its intended application, which is the *extraction of terminology*, and involve the following tasks:

- *alignment* of the texts: for the specific application purposes, alignment at sentence level has been deemed sufficient; however, the quality of the output is considered crucial; therefore, automatic processing is followed by human validation by language experts;
- external and internal *structural annotation*: the minimal requirements include segmentation at sentence level for the alignment task and metadata information that will be required for the distribution and re-use of the corpus;
- *linguistic processing*: below-Part of Speech (PoS) tagging and lemmatization is the minimum information required for the automatic term extraction task.

To ensure re-usability of the collected and processed material, compliance with the following internationally accredited standards was decided:

- the aligned material conforms to the TMX standard (Translation Memory eXchange, http://www.lisa.org/tmx/), which is XML-compliant. Being a vendor-neutral, open standard for storing and exchanging translation memories created by Computer Aided Translation (CAT) and localization tools, TMX standard was identified as a requirement for the eContent professionals. It allows easier exchange of translation memory data between tools and/or translation vendors with little or no loss of critical data during the process;
- for the external annotation, the IMDI metadata schema (IMDI, Metadata Elements for Session Descriptions, Version 3.0.4, Sept. 2003, http://www.mpi.nl/world/ISLE/schemas/schemas_frame.html) has been selected; the internal structural annotation adheres to the XCES standard, i.e. the XML version of the Corpus Encoding Standard (XCES, http://www.cs.vassar.edu/XCES/ and CES, http://www.cs.vassar.edu/CES/CES1-0.html).
- the linguistic annotation of the texts also adheres to the XCES standard, which incorporates the EAGLES guidelines for morphosyntactic annotation (http://www.ilc.cnr.it/EAGLES96/home.html).

3 Corpus construction

3.1 Text collection

In order to construct the parallel corpus, the first step consisted in the identification of potential sources, i.e. existing parallel corpora and, alternatively or additionally, textual material that could be used for the creation from scratch of the INTERA corpus.

Previous surveys (see section 2) that identify existing LRs as well as a search over the Internet attested the scarcity of available resources in the selected languages and domains, and so, the idea of re-using existing corpora was abandoned in favour of the construction of a new corpus from scratch.

The identification process of potential sources had to take into consideration the following requirements:

- to obtain texts from a variety of sources of interest to the eContent society,
- to ensure that the material was free of Intellectual Property Rights problems, either through the arrangement of specific agreements or by obtaining them from public sources.

The ideal candidates, in this respect, mainly consist of texts available over the Internet, provided by organizations/institutions that wish to make their own material available in more than one language, such as international organizations (e.g. United Nations, European Union, World Health Organization, Non-Governmental Organizations, etc.), multinational companies, companies with activities outside their own country (e.g. data describing company profiles & activities, product catalogues, etc.), public administration services (e.g. regarding bilateral agreements, regulations for immigrants, etc.), news agencies (targeting international broadcasting or for foreign language audience within their own country), official national government sites, national tourism organizations, etc. In all the above cases, the material consists of either web content per se (i.e. mainly bilingual web sites, rarely trilingual or quadrilingual) or of texts (official documents, technical reports, etc.) included in the web sites.

A more careful investigation, however, of web texts showed that although Internet is rapidly becoming multilingual, it is not yet parallel, especially as regards the languages involved in the project: most international bodies include original and translated texts but only in the more widely spoken languages. Moreover, a closer inspection of web texts that "seem" parallel, on the basis of structural similarities (e.g. similar size, paragraph segmentation, possible "anchors", such as list enumerators, etc.) showed that only sporadic parts of them were parallel. More problems arise from the fact that texts may contain large parts of foreign language material (e.g. EU regulations that include amendments to previous regulations by including the replacement text of specific paragraphs in all EU languages).

Given the above observations, cooperation with other data centers, with proven expertise in the area of LRs production for the specific project languages was sought; this would ensure content quality of the corpus, both during the selection (i.e. native speakers are better qualified to recognize true parallel material) and the encoding and validation processes, especially as regards the alignment validation and the linguistic processing. ILSP remains responsible for the construction of the Greek-English corpus, the collection and harmonization of the four subcorpora, the linguistic processing of the English texts and the addition of the IMDI metadata.

3.2 Text processing

Depending on the source that provided the original material (e.g. web site content, publishing house, translation company, etc.), different processing was required in order to arrive at the desired format adhering to the specifications set by the INTERA project; such as, indicatively:

- conversion of the original PDF/RTF/HTML etc. files into the format required by the various tools (tokenizer, aligner, tagger),
- cleanup of the texts from unwanted material (e.g. tables, figures, foreign language material, etc.)
- re-structuring of the original monolingual texts from the TMX file, when the source was the output of a Translation Memory,
- manual or semi-automatic annotation of metadata.

Each language team undertook the processing of the collected material (i.e. alignment and human validation, structural and linguistic annotation without human validation), using their own tools, thus ensuring that no time is lost over training with new tools and that the required language-dependent tools (especially taggers) used in the project are the most appropriate ones. The material to be delivered, however, at the end of all processes must be conformant to the selected standards.

The intervention of ILSP takes place only at the end of this process, with the purpose of validating the conformance of the results and of harmonizing any problematic issues. The most important point of this process is the linguistic annotation and, specifically, the harmonization of the different tagsets used. In conformance with the methodology adopted in the project, i.e. of re-using existing material, whenever possible, with the least possible interventions, so as to ensure time and cost efficiency, it was decided to re-use only existing tools for each language, without making any modifications to the tools themselves but only conversion(s) of their output. Therefore, the task of harmonizing the output with regard to the morphosyntactic tags employed by each tagger is the last stage of the procedure, where all tagsets are mapped to one, based on the EAGLES guidelines.

4 Conclusions

In this paper, we described the methodology followed in the construction of a multilingual parallel corpus; this task has been interpreted as a test application endeavor in the process of defining a business model for the LRs production. The effort was to identify gaps and shortcomings in the process usually employed by LRs producers (or users who might wish to create their own LRs) and to suggest ways of remedying them. Our findings include:

- *problems faced during the acquisition phase*: although an increasing supply of raw data (e.g. over Internet) and tools capable of exploiting this data (e.g. web crawlers that can identify and download texts in a given language) is attested, there is also a need for the enhancement of these tools with more intelligent techniques (e.g. incorporation of alignment techniques during the acquisition process in order to spot potential parallel texts, identification and mark-up of large foreign language excerpts),
- *problems faced during the processing phase*: in order to enhance the LRs production effort, the re-use of existing tools is considered crucial. It is true that an increasing number of tools are available for text processing; however, this is oriented mainly towards the major languages. Moreover, information concerning the existence, availability and operation of existing tools is not easy to locate – a gap that the other pillar of INTERA tries to remedy through the building of an integrated European Language Resources area. Additionally, tools must be enhanced with respect to two directions: improvement of the tools themselves (e.g. more robust alignment techniques) and interoperability of all relevant tools currently used at different phases of processing. The issue of interoperability is closely related with the issue of standards. The promotion and deployment of existing standards as well as the creation of new standards, when these are lacking, is important to ensure viability and re-use of LRs, given the cost of their production.

References

Gavrilidou, M., E. Desipri. 2003. Final Version of the Survey, ENABLER Deliverable 2.1.

Gavrilidou, M. E. Desipri, P. Labropoulou, S. Piperidis, N. Calzolari, M. Monachini & C. Soria. 2004. Technical specifications for the selection and encoding of multilingual resources, INTERA (Integrated European language data Repository Area), Deliverable 5.1.

IMDI, Metadata Elements for Session Descriptions, Version 3.0.4, Sept. 2003.

INTERA – eContent 2002 Integrated European languages data Repository Area, Technical Annex.

LISA. 2001. The LISA Globalization Strategies Awareness Survey.

LISA/AIIM. 2001. The Black Hole in the Internet: LISA/AIIM Globalization Survey.

LISA/OSCAR. 2003. Translation Memory Survey.

Maegaard, B., K. Choukri, V. Mapelli, M. Nikkhou & C. Povlsen. 2003. Language resources-Industrial needs, ENABLER Deliverable 4.2.