A web-enabled and speech-enhanced parallel corpus of Greek - Bulgarian cultural texts

Voula Giouli
Institute for Language & Speech Processing Athens, Greece
voula@ilsp.gr

Nikos Glaros
Institute for Language & Speech Processing Athens, Greece
nglaros@ilsp.gr

Kiril Simov
Institute for Parallel Processing, BAS, Sofia, Bulgaria
kivs@bultreebank.or

Petya Osenova
Institute for Parallel Processing, BAS, Sofia, Bulgaria
petya@bultreebank.org

Abstract

This paper reports on completed work carried out in the framework of an EU-funded project aimed at (a) developing a bilingual collection of cultural texts in Greek and Bulgarian, (b) creating a number of accompanying resources that will facilitate study of the primary texts across languages, and (c) integrating a system which aims to provide web-enabled and speech-enhanced access to digitized bilingual Cultural Heritage resources. This simple user interface, which incorporates advanced search mechanisms, also offers innovative accessibility for visually impaired Greek and Bulgarian users. The rationale behind the work (and the relative resource) was to promote the comparative study of the cultural heritage of the two countries.

1 Introduction

The document describes a bilingual Greek (EL) and Bulgarian (BG) collection of literary and folklore texts along with the metadata that were deemed necessary for the efficient management and retrieval of the textual data. Section 2 outlines the project aims that guided selection and annotation of the texts, whereas Section 3 presents the primary data that comprise the bilingual textual collection and the methodology adopted for collecting them. Section 4 elaborates on the metadata scheme that has been implemented to describe the primary data and the linguistic annotation tailored to facilitate search and retrieval at the document, phrase or word level. This scheme is compliant to widely accepted standards so as to ensure reusability of the resource at hand. Section 5 presents the Language Technologies (LT) deployed in the project elaborating on the Greek and the Bulgarian text processing tools, and discusses the LT methods that have been (a) exploited in the course of the project to facilitate the web-interface construction and (b) integrated in the search and retrieval mechanisms to improve the system performance. Finally, Section 6 describes the main components of the web interface and the way various features are exploited to facilitate users' access to the data. In the last section, we present conclusions and future work.

2 Project description

The project aims at highlighting cultural resources that, as of yet, remain non-exploited to their greatest extent, and at creating the necessary infrastructure with the support of LT with a view to promoting the study of cultural heritage of the eligible neighboring areas and raising awareness about their common cultural identity. To serve these objectives, the project had a concrete target, that is, the creation of a textual collection and of accompanying material that would be appropriate for the promotion and study of the cultural heritage of the neighboring areas in Greece and Bulgaria (Thrace and the neighboring Smolyan, Blagoevgrad, Kardjali, Khaskovo areas), the focus being on literature, folklore and language. To this end, the main activities within the project life-cycle were to:

- record and roadmap the literary production of the afore mentioned areas spanning from the 19th century till the present days along with written records on folk culture and folktales from the eligible areas. These should form a pool of candidate texts from which

Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education –LaTeCH – SHELT&R 2009, pages 35–42,
Athens, Greece, 30 March 2009. ©2009 Association for Computational Linguistics

the most appropriate for the project objectives could be selected;

- record and roadmap existing translations of literary works in both languages to serve for the creation of the parallel corpus;
- select textual material representative of the two cultures, and thus, suitable for their comparative study;
- digitize the selected (printed) material to a format suitable for long-term preservation;
- collect meta-texts relevant to the selected literary and folklore texts, that is, texts about the literary works, biographies of the selected authors, criticism, etc.; these comprise part of the accompanying material
- document the data with any information deemed necessary for its preservation and exploitation, catering for their interrelation so as to highlight their common features and allow unified access to the whole set along text types / genres and languages;
- extract bilingual glossaries from the primary collection of literary and folklore texts also accounted for as accompanying material; the project caters for the extraction of EL and BG terms and names of Persons and Locations and their translation equivalents in the other language;
- make the primary resource along with the accompanying material (meta-texts and glossaries) publicly available over the internet to all interested parties, ranging from the research community to laypersons, school students and people interested in finding out more about the particular areas;
- facilitate access to the material that wouldn’t be hampered by users’ computer literacy and/or language barriers. To cater for the latter, the web interface would be as simple as possible – yet functional – and the data should be available in both languages (Greek and Bulgarian) plus in English.

3 The bilingual Greek – Bulgarian Cultural Corpus

Along with the aforementioned lines, the collection comprises parallel EL – BG literary and folklore texts. The main specifications for the Greek - Bulgarian Cultural Corpus (GBCC) creation were:

- to build a bilingual resource that could be used as a means to study cultural similarities and/or differences between the neighboring areas of Greece and Bulgaria the focus being on literature, folklore and folktales;
- to provide a representative sample of (a) literature written by authors from Thrace -that is from the entire area of Thrace- or about Thrace, spanning between the 19th century - today, (b) folklore texts about Thrace, that would normally reflect cultural as well as linguistic elements either shared by the two people or unique to each culture, and (c) folktales and legends from Thrace, the latter being the intermediate between literature and folklore.

In order to gather the candidate texts and authors for such a collection we exploited both printed and digitized sources, i.e., (on-line and printed) anthologies of Bulgarian, Greek or Balkan literature, digital archives, web resources and library material. The outcome of this extensive research was a wealth of literary works including titles by the most prominent authors in Bulgaria and Greece. The selection of the authors, who would finally participate in GBCC, was based on the following criteria: (a) author's impact to Greek or Bulgarian literature respectively; and (b) author's contribution to his county's folk study or other major sectors such as journalism and education.

Additionally, to ensure corpus “representativeness” to some extend, we tried to include the full range of the literary texts (poetry, fiction, short stories) and in proportion to the literary production with respect to the parameters of place, time and author. To this end, we think we have avoided biases and the corpus models all language varieties spoken in the areas and at different periods.

Moreover, the "inner" content characteristics of texts were used as the basic criteria for text selection. To this end, we chose texts which demonstrate the two people's cultural similarities and affinity along with each author's most important and representative works. Beyond the above, the availability of a translation in the other language and IPR issues also influenced text selection.

The collection of the primary data currently comprises of (135) literary works, (70) BG (Bulgarian) and 65 EL (Greek). Moreover, (30) BG folk texts and 30 EL folk texts along with (25) BG folktales and 31 EL folktales were added in order to build a corpus as balanced as possible and representative of each country's culture. In terms of tokens, the corpus amounts to 700,000 in total (circa 350,000 tokens per language): the literature part is about 550,000 tokens, whereas, the folklore and legend sub-corpus is about 150,000 tokens.

Moreover, to cater for the project requirement that the corpus should be bilingual, available translations of the primary EL – BG literary works were also selected to form the parallel literary corpus. Additionally, an extensive translation work was also carried out by specialized translators where applicable (folklore texts and folktales).

The collection covers EL and BG literary production dating from the 19th century till the present day, and also texts (both literary or folklore) that are written in the dialect(s) used in the eligible areas. This, in effect, is reflected in the language varieties represented in the textual collection that range from contemporary to non-contemporary, and from normal to dialectical or even mixed language.

Finally, the collection of primary data was also coupled with accompanying material (content metadata) for each literary work (literary criticism) and for each author (biographical information, list of works, etc.). Along with all the above, texts about the common cultural elements were also included.

4 Corpus Annotation

After text selection, digitization and extended manual validation (where appropriate) were performed. Normalization of the primary data was kept to a minimum so as to cater, for example, for the conversion from the Greek polytonic to the monotonic encoding system. Furthermore, to ensure efficient content handling and retrieval and also to facilitate access to the resource at hand via the platform that has been developed, metadata descriptions and linguistic annotations were added across two pillars: (a) indexing and retrieval, and (b) further facilitating the comparative study of textual data. To this end, metadata descriptions and linguistic annotations compliant with internationally accepted standards were added to the raw material. The metadata scheme deployed in this project is compliant with internationally accredited standards with certain modifications that cater for the peculiarities of the data.

More specifically, the metadata scheme implemented in this project builds on XCES, 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), which has been proposed by EAGLES (http://www.ilc.cnr.it/EAGLES96/home.html) and is compliant with the specifications of the Text Encoding Initiative (http://www.tei-c.org, Text Encoding Initiative (TEI Guidelines for Electronic Text Encoding and Interchange). From the total number of elements proposed by these guidelines, the annotation of the parallel corpus at hand has been restricted to the recognition of structural units at the sentence level, which is the minimum level required for the alignment and term extraction processes. That means that the requirements of CES Level 1 conformance are met; as regards CES Level 2 the requirements (but not the recommendations) are also met, and from CES Level 3 requirements, annotation for sentence boundaries is met.

Additionally, metadata elements have been deployed which encode information necessary for text indexing with respect to text title, author, publisher, publication date, etc. (bibliographical information) and for the classification of each text according to text type/genre and topic, the latter being applicable to folklore texts and folk tales. Classification of folklore texts is based on the widely accepted Aarne-Thompson classification system (Aarne, 1961).

To this end, to assure documentation completeness, and facilitate the inter-relation among primary data and the accompanying material (biographies, criticism, etc) the documentation scheme has been extended accordingly. The aforementioned metadata descriptions are kept separately from the data in an xml header that is to be deployed by the web interface for search and retrieval purposes.

The external structural annotation (including text classification) of the corpus also adheres to the IMDI metadata scheme (IMDI, Metadata Elements for Session Descriptions, Version 3.0.4, Sept. 2003). Adaptations proposed specifically concerning Written Language Resources have been taken into account. IMDI metadata elements for catalogue descriptions (IMDI, Metadata Elements for Catalogue Descriptions, Version 2.1, June 2001) were also taken into account to render the corpus compatible with existing formalisms (ELRA, and LDC). This type of metadata descriptions was added manually to the texts.

To further enhance the capabilities/functionalities of the final application, rendering, thus the collection a useful resource to prospective users and researchers, further annotations at various levels of linguistic analysis were integrated across two pillars: (a) efficient indexing and retrieval; and (b) further facilitating the comparative study of textual data by means of bilingual glossaries which were constructed semi-automatically, and via the visualization of aligned parallel texts.

Text processing at the monolingual level comprises the following procedures: (a) handling and tokenization, (b) Part-of-Speech (POS) tagging and lemmatization, (c) surface syntactic analysis, (d) indexing with terms/keywords and phrases/Named Entities (NEs) pertaining to the types Location (LOC) and Person (PER).

Annotations at these levels were added semi-automatically, by deploying existing generic Natural Language Processing (NLP) tools that were developed for the languages at hand, whereas extensive and intensive validations were performed via several ways. Indeed, although the tools deployed have reported to achieve high accuracy rates in the domains/genres they were intended for, the specific nature of the data led to a significant reduction. To this end, half of the annotations were checked manually. After the identification of the errors in this part of the corpus, we have performed a manual check in the second part of the corpus only for these cases which were recognized as errors during the validation of the first part. For some of the cases relevant constraints in the systems were written, which automatically find places where some rules were not met. Tools customization was also performed by adding new rules applicable for the language varieties to be handled, and also by extending/modifying the resources used (word and name lists, etc.).

Finally, alignment of parallel texts (primary source documents and their translations) has also been performed at both sentence and phrase level. As expected, poems posited the major difficulties due the fuzziness in identifying sentence boundaries, and alignments at the phrase level were favored instead.

5 Language Technologies

In what follows the Greek and Bulgarian Text Processing Components will be described.

5.1 The Greek pipe-line

In the case of the Greek data, text processing was applied via an existing pipeline of shallow processing tools for the Greek language. These include:

- Handling and tokenization; following common practice, the Greek tokenizer makes use of a set of regular expressions, coupled with precompiled lists of abbreviations, and a set of simple heuristics (Papageorgiou et al., 2002) for the recognition of word and sentence boundaries, abbreviations, digits, and simple dates.
- POS-tagging and lemmatization; a tagger that is based on Brill's TBL architecture (Brill, 1997), modified to address peculiarities of the Greek language (Papageorgiou et al., 2000) was used in order to assign morphosyntactic information to tokenized words. Furthermore, the tagger uses a PAROLE-compliant tagset of 584 different part-of-speech tags. Following POS tagging, lemmas are retrieved from a Greek morphological lexicon.
- Surface syntactic analysis; the Greek chunker is based on a grammar of 186 rules (Boutsis et al., 2000) developed for the automatic recognition of non-recursive phrasal categories: adjectives, adverbs, prepositional phrases, nouns, verbs (chunks) (Papageorgiou et al., 2002).
- Term extraction; a Greek Term Extractor was used for spotting terms and idiomatic words (Georgantopoulos, Piperidis, 2000). Term Extractor's method proceeds in three pipelined stages: (a) morphosyntactic annotation of the domain corpus, (b) corpus parsing based on a pattern grammar endowed with regular expressions and feature-structure unification, and (c) lemmatization. Candidate terms are then statistically evaluated with an aim to skim valid domain terms and lessen the overgeneration effect caused by pattern grammars (hybrid methodology).

Named Entity Recognition was then performed using MENER (Maximum Entropy Named Entity Recognizer), a system compatible with the ACE (Automatic Content Extraction) scheme, catering for the recognition and classification of the following types of NEs: person (PER), organization (ORG), location (LOC) and geopolitical entity (GPE) (Giouli et al., 2006).

5.2 Bulgarian Tools

In the processing of the Bulgarian part of the corpus we have been using generic language technology tools developed for Bulgarian. Here is the list of tools that we have used. They are implemented within the CLaRK System (Simov et al. 2001) via:

Tokenization, Morphosyntactic tagging, Lemmatization; Tokenization is implemented as a hierarchy of tokenizers within the CLaRK system. Morphosyntactic tagging is done on the basis a morphological lexicon which covers the grammatical information of about 100 000 lexemes (1 600 000 word forms); a gazetteers of about 25000 names and 1500 abbreviations. We are using the BulTreeBank tagset, which is a more specialized version of Multext-east tagset. The disambiguation is done in two steps. Initially, a rule-based module solves the sure cases for which manual rules can be written. Then, for the next step, a neural-network-based disambiguator is being exploited (Simov and Osenova 2001). Lemmatization is implemented as rules which convert each word form in the lemma. The rules are assigned to the word forms in the lexicon. This ensures very high level of accuracy.

Partial Grammars have also been constructed for *Sentence splitting, Named-entity recognition*, and *Chunking*.

5.3 Alignments

To facilitate the comparative study of parallel documents, source texts were automatically aligned with their translations. Alignments at the sentence level were performed semi-automatically by means of the ILSP Aligner, which is a language independent tool that uses surface linguistic information coupled with information about possible unit delimiters depending on the level at which the alignment is sought. The resulting translation equivalents were stored in files conformant to the internationally accredited TMX standard (Translation Memory eXchange, http://www.lisa.org/tmx/), which is XML-compliant, vendor-neutral open standard for storing and exchanging translation memories created by Computer Aided Translation (CAT) and localization tools.

Moreover, terms pertaining to the folklore domain as well as names of Persons and Locations identified in the EL - BG parallel texts were semi-automatically aligned. The outcome of the process of text alignment at below the sentence level was then validated manually.

5.4 Tools Customization and metadata harmonization

As it has already been stated, the tools that were deployed for the linguistic processing are generic ones that were initially developed for different text types/genres. Moreover, the data at hand posed another difficulty that is, coping with older/obsolete language usage. In fact, some of the literary works were written in the 19th century or the beginning of 20th century, and their language reflects the writing standards of the corresponding period.

Therefore, as it was expected, the overall performance of the afore-mentioned tools was lower than the one reported for the texts these tools were initially trained for.

To this end, performance at POS-tagging level dropped from 97% to 77% for the Greek data since no normalization of the primary data was performed. On the other hand, the BG morphological analyzer coverage, whose benchmark performance is 96% dropped to 92 % on poems and folktales and to 94% on literary texts and legends. The reason was that the language of processed literary texts and legends came normalized from the sources, while the poems and folktales kept some percentage of archaic or dialect words. Thus, additionally to the guesser, a post POS processing was performed on the unknown words. Moreover, the accuracy of the neural network disambiguator and the rule-based one was 97 %. i.e. the same as for other applications. Processing at the levels of chunks and NEs were even lower. Within the project we had to tune the tools to the specific language types, such as diachronically remote texts and domain specific texts (folklore). Also, some words with higher distribution in the target regions appear in some of the works. In order to deal with them we had to extend the used lexicons, to create a guesser for the unknown words and add new rules to the chunk grammar to handle some specific word order within the texts.

Additionally, the deployment of tools that are specific to each language and compatible with completely distinct annotation standards brought about the issue of metadata harmonization. To this end, although the Greek tools were developed to confront to the afore-mentioned annotation standards, this was not the case for Bulgarian. The first encoding scheme followed the BulTreeBank morphological and chunk annotation scheme. Afterwards, the information was transferred into the project scheme in order to be consistent with the Greek data and applicable for web representation. As a result, the morphosyntactic features of the BG tagset, which is a more specialized version of the

Multext-East tagset were mapped onto the relative PAROLE tags.

6 The web interface

All the data collected (being the primary literary or folklore texts or meta-documents, etc.) along with their translations, the multi-layered annotations, and the resulting glossaries were integrated in a database platform that was developed to serve as a content management system. Being the backbone of that platform, the metadata material facilitates the interlinking of similar documents, and the access to the primary data via the web. To this end, a specially designed web site was developed to satisfy the needs of end-users (the general public and the special groups of researchers and other scientists). The website features a trilingual interface (Greek, Bulgarian, English) as well as advanced search and retrieval mechanisms on the entire bilingual content or a user-specified part of it. The users can perform combined searches by author name, title, genre, etc. Furthermore, they can search for single keywords/wordforms or for two wordforms that can be a user-specified number of words apart from each other. Searches by lemma and/or by phrase have been also implemented. The latter rely on a matcher, which tries to link the query word(s) with the stored lemmas/wordforms. Additionally, a stemmer for Greek and Bulgarian has been used for the online stemming of queries, which will then be matched with the already stemmed corpus. When all the above fails, fuzzy matching techniques are being employed, facilitating, thus, effective query expansion functionality. Finally, apart from wordforms and lemmas, the collection can also be queried for morphosyntactic tags or any combination thereof; results, then, come in the form of concordances and statistics (frequency information), hence the relative document(s) can also be retrieved. Moreover, users can search the whole corpus or define a sub-corpus based on the classification and annotation parameters accompanying each text, thus, creating sub-corpora of a specific author, or belonging to a specific genre, text type, domain, time period, etc.

In addition, the web interface lets the users to simultaneously view on screen both Greek and Bulgarian texts, aligned and in parallel,, so that to become acquainted with the comparative aspects of the two languages or perform specific linguistic, lexicographic or translation tasks. Alternatively, the user can consult the bilingual glossary of terms and the aligned list of NEs. The latter is often very interesting, especially with respect to Location entities, since transliteration is usually non-adequate.

The design of the web interface effectively blends simplicity and advanced functionality so that to fully support the intended usage scenarios (comparative study of literary and folklore texts equally by specialists, laymen or students, language and/or literary teaching and learning, lexicographic projects, etc.). Finally, the web interface has been enhanced by integrating last generation of synthetic speech technology for both Greek and Bulgarian. This speech-enhanced user interface (S. Raptis et al, 2005), offers innovative web accessibility for blind and vision impaired Greek and Bulgarian users as well as for other users who use speech as their preferable modality to information access. The key-feature of this web-speech technology is that it lets users to interact with the underlying system; so that they can hear only the portions of a specific web page they are interested in, being able at the same time to navigate through the entire web site and visit only the web pages of their choice.

7 Conclusions and future work

We have described work targeted at the promotion and study of the cultural heritage of the cross-border regions of Greece – Bulgaria, the focus been on literature, folklore and language of the two people, by means of modern and technologically advanced platforms. To this end, a digital collection of literary and folklore texts has been compiled along with accompanying material selected from various (online and printed sources), which is integrated into a platform with advanced search and retrieval mechanisms.

However, the cultural value of the bilingual cultural Greek-Bulgarian corpus goes beyond the border areas that it was intended for, because it shows the similarities and the differences between the two neighboring countries. More specifically, it can be used for supporting the acquisition of the other language in both countries. Also, it can be explored for comparing the cultural and social attitudes in diachronic depth and genre variety. Apart from the usages from a humanities point of view, the corpus can become a good base for testing taggers, parsers and aligners. It would especially challenge the processing of the regional dialects, the language of poems, and the language of non-contemporary works.

Future work is being envisaged in the following directions: extending the corpus with more texts, and respectively the glossaries – with more terms, adding more layers of linguistic analysis (predicate-argument structure, etc.), and further enhance search and retrieval with the construction and deployment of an applicable thesaurus.

Acknowledgments

We would like to thank the anonymous reviewers for useful suggestions and comments. Most of the work presented in this paper was done in the framework of a project that was funded under the Community Initiative Programme INTERREG III A / PHARE CBC Greece – Bulgaria. The project was implemented by the Institute for Language and Speech Processing (ILSP, www.ilsp.gr) and a group of academics and researchers from the Sofia University St. Kliment Ohridski (www.uni-sofia.bg).

References

Antti Aarne. 1961. *The Types of the Folktale: A Classification and Bibliography. Translated and Enlarged by Stith Thompson.* 2nd rev. ed. Helsinki: Suomalainen Tiedeakatemia / FF Communications.

Sotiris Boutsis, Prokopis Prokopidis, Voula Giouli and Stelios Piperidis. 2000. *A Robust Parser for Unrestricted Greek Tex*. In Proceedings of the 2nd Language and Resources Evaluation Conference, 467-473, Athens, Greece.

Michel Généreux. 2007. Cultural Heritage Digital Resources: From Extraction to Querying, Language Technology for Cultural Heritage Data (LaTeCH 2007), Workshop at ACL 2007, June 23rd–30th 2007, Prague, Czech Republic.

Byron Georgantopoulos and Stelios Piperidis, 2000. *Term-based Identification of Sentences for Text Summarization*. In Proceedings of LREC2000

Voula Giouli, Alexis Konstandinidis, Elina Desypri, Harris Papageorgiou. 2006. *Multi-domain Multilingual Named Entity Recognition: Revisiting & Grounding the resources issue*. In Proceedings of LREC 2006.

IMDI, Metadata Elements for Catalogue Descriptions, Version 2.1, June 2001

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

Harris Papageorgiou, L. Cranias, Stelios Piperidis1994. *Automatic alignment in parallel corpora.* In Proceedings of ACL 1994.

Harris Papageorgiou, Prokopis Prokopidis, Voula Giouli, Iasonas Demiros, Alexis Konstantinidis, and Stelios Piperidis. 2002. *Multi-level XML-based Corpus Annotation*. Proceedings of the 3nd Language and Resources Evaluation Conference.

Harris Papageorgiou, Prokopis Prokopidis, Voula Giouli, and Stelios Piperidis. 2000. *A Unified POS Tagging Architecture and its Application to Greek*. In Proceedings of the 2nd Language and Resources Evaluation Conference, Athens, Greece, pp 1455-1462.

Stelios Piperidis. 1995. *Interactive corpus based translation drafting tool.* In ASLIB Proceedings 47(3), March 1995.

Spyros Raptis, I. Spais and P. Tsiakoulis. 2005. *A Tool for Enhancing Web Accessibility: Synthetic Speech and Content Restructuring"*. In Proc. HCII 2005: 11th International Conference on Human-Computer Interaction, 22-27 July, Las Vegas, Nevada, USA.

Kiril Simov, Z. Peev, M. Kouylekov, A. Simov, M. Dimitrov, and A. Kiryakov. 2001. *CLaRK - an XML-based System for Corpora Development.* Corpus Linguistics 2001 Conference. pp 558-560.

Kiril Simov, and Petya Osenova. *A Hybrid System for MorphoSyntactic Disambiguation in Bulgarian*. In: Proc. of the RANLP 2001 Conference, Tzigov Chark, Bulgaria, 5-7 September 2001. pages 288-290.

René Witte, Thomas Gitzinger, Thomas Kappler, and Ralf Krestel. 2008. A Semantic Wiki Approach to Cultural Heritage Data Management. Language Technology for Cultural Heritage Data (LaTeCH 2008), Workshop at LREC 2008, June 1st, 2008, Marrakech, Morocco.