Language Technologies customized for processing Greek textual Cultural Heritage data

Voula Giouli Evi Marzelou Prokopis Prokopidis Maria Zourari
Institute for Language and Speech Processing

1 Introduction

This paper reports on work aimed at (a) developing an application tailored to integrate and highlight textual cultural resources that, as of yet, remain under-exploited, and (b) creating the necessary infrastructure with the support and customization of Language Technologies (LT). The ultimate goal was to promote the study of cultural heritage of Greece and Bulgaria (the focus being on the neighboring areas) and to raise awareness about their common cultural identity, emphasizing on literature, folklore and language. To this end, a bilingual collection of literary and folklore texts in Greek and Bulgarian—called hereafter Greek-Bulgarian Cultural Corpus (GBCC)—was developed along with a number of accompanying resources that were extracted semi-automatically from the textual data. More precisely, we elaborate on the Greek counterpart of the textual collection and the processing applied thereof.

In the paper, we will first present the rationale behind this endeavor, i.e., developing an infrastructure with the help of HLT (section 2), and the specific targets of the project at hand (section 3). We will then describe the collection in terms of the textual data, the metadata added manually and the processing applied via an existing pipeline of shallow processing tools elaborating on the Greek text processing tools that are integrated in the cross-lingual search and retrieval mechanisms of the fully-functional platform (section 4). Ongoing work towards customizing the POS-tagger with respect to the language varieties covered by the data is presented in section 5. Conclusions and future work is finally reported in section 6.

2 HLT for cultural heritage content

Cultural heritage, that is, the legacy of physical or spiritual artifacts and intangible attributes of a group/society over a certain period in time, has become the centre of attention to a diverse audience. This growing interest has led to an increase in large-scale digitization efforts and the development of cultural collections and/or digital libraries targeted to wide-spread and complex audiences with disparate or even competing interests. And, although there has been a long tradition in the digitization and manual documentation of cultural heritage data, yet the need for indexing and retrieval of cultural content that goes beyond mere bibliographic information has only recently been recognized. Annotating, therefore, documents with useful metadata such as person or location names, indexes of events, etc. is currently considered as important for the effective management and retrieval of cultural content. Never the less, adding metadata manually is difficult, laborious and costly a task. And, while fully-automatic solutions towards indexing cultural content have been proved neither feasible nor practical, HLT is currently used as a valuable aid towards bootstrapping such laborious and expensive undertakings (Bontcheva et al. 2002, Nissim et al. 2004, Borin et al., 2007, Borin et al. 2010, etc.). A number of initiatives acknowledge that language technology can offer new ways of accessing and visualizing cultural content. This is also true for textual collections or archives, where automatic content annotation and indexing not only facilitates the retrieval processes, but can also give rise to new types of scholarship.

Indeed, recent advances in HLT in terms of both accuracy and robustness of the software solutions available for a wide range of languages have opened new perspectives in this respect. These technologies include not only baseline tools for Part-Of-Speech tagging and lemmatization, but also modules such as, for example, Named Entity Recognition, Information Extraction and Retrieval, Summarization, Event Annotation, Emotion Recognition, etc. However, although HLT seems to be mature enough to cope with a number of applications, existing robust systems have been developed with specific applications in mind and even focus on specialized text types and narrow sub-domains (i.e., finance, terrorist attacks, etc.) With this respect, one of the most central problems encountered is that of adapting generic modules to new genres, text types or even languages and language varieties, placed on the temporal or spatial axis (older forms of a given language, or dialects, respectively).

All the above, that is, the growing tendency for creating cultural heritage textual collections coupled with metadata for a variety of languages, and the urgent need for portability and customization generic HLT tools, has brought about the need for a basic research infrastructure for language technology. At the core of this infrastructure lies the so-called BLARK – Basic Language Resource Kit. This notion is used to refer to a core set of language resources and LT tools deemed essential both to basic research in language technology and to the development of HLT applications for a particular language (i.e., linguistically annotated text corpora, lexical resources, tools for linguistic annotation of tools, etc). A BLARK usually refers to modern standard languages, which are topic- and genre-neutral. However, in the context of texts written in the non-standard language variety, researchers have been engaged in the development of a suitable BLARK (Borin et al. 2010). Indeed, this need is increasingly recognized by the language resource community and research funding agencies alike, and to this respect, the work presented here was perceived as a contribution to the creation of a diachronic BLARK for the Greek language.

3 Project description

The main activities within the project life-cycle can be outlined as follows: (1) record and roadmap the literary production of the afore mentioned areas spanning from the 19th century until 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 the most appropriate for the project objectives could be selected; (2) record and roadmap existing translations of literary works in both languages to serve for the creation of the parallel corpus; (3) select textual material representative of the two cultures, and thus, suitable for their comparative study; (4) digitize the selected (printed) material to a format suitable for long-term preservation; (5) 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 (6) 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; (7) 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; (8) 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; (9) create an infrastructure to facilitate access to the material that wouldn't be hampered by users' computer literacy and/or language barriers.

4 Collection description

4.1 The textual material

The corpus comprises three text types: (a) literary works either written by authors from Thrace and the neighboring Bulgarian areas or with a story situated in Thrace; (b) folklore texts, i.e., those depicting a wide range of aspects of human activity such as traditions, customs, practices, spiritual beliefs and other aspects of everyday life in the eligible areas; and (c) folktales and legends from the entire area of Thrace.

In order to gather the candidate texts and authors for such a collection we exploited both printed and digitized sources (i.e., 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 in Bulgarian (BG) and 65 in Greek (EL)). 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.

As it has already be mentioned, to cater for the project requirement that the corpus should be bilingual, 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 literary production in Greece and Bulgaria dating from the 19th century until 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, and also some historical texts depicting the origins of places, people, major historical events, etc.

4.2 The metadata schema

After text selection, digitization and extended manual validation where performed where appropriate. 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, www.cs.vassar.edu/XCES/ and CES, www.cs.vassar.edu/CES/ CES1-0.html), which has been proposed by EAGLES (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 and phrase level, which since these were deemed necessary for the alignment and term extraction processes.

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). Additionally, to ensure documentation completeness, and facilitate the inter-relation among primary data and the accompanying material (biographies, criticism, etc) the documentation scheme has been extended accordingly so as to include these elements. Information regarding text type/genre and topic was also added. Folklore texts have been classified following the Library of Congress Classification scheme (http://www.loc.gov/catdir/cpso/lcco/), whereas folktales are categorized on the basis of the Aarne-Thompson classification system (Aarne 1961). Finally, information on certain characteristics of the texts, such as language variety (contemporary/non-contemporary/idiomatic), etc. was also added to the metadata descriptions.

The aforementioned metadata descriptions are kept separately from the primary 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). 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 standards (ELRA, and LDC). This type of metadata descriptions was added manually to the texts.

4.3 Levels of Linguistic Analysis

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), and (e) alignment of parallel texts at sentence and phrase levels. Processing on the Greek textual data was applied via an existing pipeline of shallow processing tools for the Greek language.

At the first stage, handling and tokenization was performed using a Greek tokenizer that employs 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.

At the next stage, POS-tagging and lemmatization were performed in order to assign morphosyntactic characteristics and lemma information to every token in the text. To accomplish this task, we used the POS-tagger developed in-house that is based on Brill's TBL architecture (Brill 1997), which had been trained on Greek textual data from various sources (newspapers, internet, etc.). The tagger uses a PAROLE-compliant tagset of 584 different part-of-speech tags (Papageorgiou et al. 2000) and captures the peculiarities of the Greek language. Following POS tagging, lemmas were then retrieved from a Greek morphological lexicon and assigned to every word form.

Surface syntactic analysis was performed by the Greek rule-based chunker developed for the automatic recognition of non-recursive phrasal categories: adjectives, adverbs, prepositional phrases, nouns, verbs (chunks) (Papageorgiou et al. 2002).

At the next stage, a Greek Term Extractor tailored to spotting terms and idiomatic words (Georgantopoulos&Piperidis 2000) was employed. Term Extractor functions 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 over-generation effect caused by pattern grammars.

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). For the purposes of the current project, only NEs pertaining to the types Location (LOC) and Person (PER) were retained since they seemed more appropriate for the data at hand and the use searches that would be of interest.

Finally, source texts were automatically aligned with their translations in order to facilitate reading comprehension for speakers of both languages. 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.

The outcome of the process of text alignment at below the sentence level was validated manually.

4.4 Website functionalities

As it has already been pointed out, the ultimate goal of the project was to create a set of language resources along with an infrastructure targeted to a wide and rather diverse audience. The application is aimed to serve as a teaching aid either in the domain of literature and folklore, or even in language teaching and learning. A more ambitious target of the project was to familiarize scholars in the humanities with applications assisting their research, and to raise awareness amongst scholars and researchers in the humanities with respect to the digital resources and advanced applications capabilities.

To this end, a website was developed that features a trilingual interface (Greek, Bulgarian, and English) as well as advanced search and retrieval mechanisms. 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.

The collection can be navigated by language, genre, text type, and/or author. Folklore texts and folktales can also be browsed by the category they fall in. The metadata material also facilitates the interlinking of similar documents (literary works with biographies or criticisms, etc). In addition, a visualization tool integrated in the platform, allows users to simultaneously view on screen the aligned Greek and Bulgarian texts in parallel, facilitating, thus, comprehension of texts in both languages. Alternatively, end-users are provided with an on-line bilingual glossary of terms and place names.

Additionally, users can perform simple or advanced searches for texts or words/lemmas, and documents can be retrieved not only on the basis of bibliographic information (author name, title, genre, etc) but also on the basis of their content. This advanced search mechanism also supports morphologically aware search and retrieval.

Furthermore, linguistically-oriented searches are also allowed for single keywords/wordforms or for combinations thereof (i.e., consecutive keywords/wordforms or ones that are separated by one or more words), and also searches for lemma and/or phrase. 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 on-line stemming of queries, which will then be matched with the already stemmed corpus. When all the above fail, fuzzy matching techniques are being employed facilitating, thus, an 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, on the basis of the metadata specifying bibliographic information, users can create sub-corpora (of a specific author, or belonging to a specific genre, text type, domain, time period, etc.) and conduct their searches thereof.

Finally, a tool capable of constructing profiles of words or word classes (for example, verbs, adjectives, adverbs, etc.) integrates statistical data and interesting searches might be elaborated on the whole corpus or a sub-corpus revealing word usage or shifts in sense according to the genre a text pertains to, or even senses within the axis of diachrony, etc.

The design of the web interface blends simplicity and advanced functionality so as to support the intended usage scenarios (i.e., the comparative study of literary and folklore texts, language and/or literary teaching and learning, lexicographic projects, studies over the style of a certain author, etc.).

5 Customization of generic NLP Tools: the POS-tagger

The project posed two main challenges: (a) processing literary texts, and (b) processing texts written in the non-standard language variety of the late 19th and early 20th centuries or the language variety spoken in the region. As expected, although the tools deployed have reported to achieve high accuracy rates in a number of applications, the specific nature of the data led to a significant reduction. Tuning of the tools to the specific language types/varieties was, therefore, considered of prominent importance. Being a baseline tool and one of the first ones in the pipe-line, the priority was given to the POS-tagger and to tackling the non-contemporary texts.

5.1 The non-contemporary corpus

A test-bed was constructed for the formal evaluation of our generic POS-tagger and for studying all the problematic cases that should be taken into account. All texts assigned the value “non-contemporary” for the attribute “language-variety” were automatically selected from the entire corpus. The resulting testbed, hence called non-contemporary corpus amounts to c. 98k words, and comprises mainly folklore and historical texts.

5.2 Towards tagger customization

Annotations applied to the texts were automatically checked manually by expert linguists using a graphical user interface suitable for manual annotation, verification and correction on the processed texts. It should be noted, however, that this was not a trivial task and posed many difficulties even to trained annotators, due to the fact that we had to cope with a number of phenomena not present in Modern Greek. After multiple passes over the data and the identification of the errors in the corpus, the following preliminary actions were taken towards the customization of the POS-tagger:

- Tagset extension (in compliance to the PAROLE specifications) so as to cater for the morpho-syntactic characteristics of the “katharevousa” (participles, infinitives, the dative case for articles, nouns, adjectives, pronouns, etc., or the morphologically distinct subjunctive mood for verbs, etc.).
- Revision of the tagging specifications so as to capture the peculiarities of the language variety at hand—“katharevousa”. Adherence to this set of instructions to human annotators ensures a high rate of inter-annotator agreement.
- Finalization of the manual annotation was then performed on the basis of the specifications set.
- Enrichment and revision of the lexicons employed by the POS-tagger, i.e., closed categories wordlists, such as to include pronouns, adverbs, prepositions of the “katharevousa”, etc. These were extracted from the validated material and further enhanced with entries from various sources (i.e., grammars, etc.)
- Word lists were further enriched with ambiguous words and wordforms that are specific to the language variety at hand.

All the afore-mentioned lexical resources (lexicons, wordlists) will be added to the resources employed by the tagger and formal validation performed.

6 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. The cultural value of this resource goes beyond the border areas that it was intended for, since it can be used for a wide range of purposes and by a diverse target group. Apart from the usages from a humanities point of view, the corpus can become a good base for developing and testing taggers, parsers and aligners. It would especially challenge the processing of the regional dialects (Greek and Bulgarian), 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, adding new layers of linguistic analysis (predicate-argument structure, etc.), and further enhance search and retrieval with the construction and deployment of an applicable thesaurus.

Acknowledgments

Most of the work presented in this paper was done in the framework of the project “***Cultural Parallels*: Study and Promotion of the Cultural Heritage of the Neighbouring Areas in Greece and Bulgaria through Internet Technology**” that was funded under the Initiative 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

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

Bontcheva, K., D. Maynard, H. Cunningham, and H. Saggion. 2002. Using Human Language Technology for Automatic Annotation and Indexing of Digital Library Content. Lecture Notes In Computer Science, Vol. 2458. In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, 613–625.

Borin, L., D. Kokkinakis, and L. J. Olsson. 2007. Naming the past: Named entity and animacy recognition in the 19th century Swedish literature. In *Proceedings of the ACL Workshop: Language Technology for Cultural Heritage Data (LaTeCH.)*, 1–8. Prague: ACL.

Borin, L., M. Forsberg, and D. Kokkinakis. 2010. Diabase: Towards a diachronic BLARK in support of historical studies. In *Proceedings of LREC 2010*.

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

Giouli, V., A. Konstandinidis, E. Desypri, and H. Papageorgiou. 2006. Multi-domain Multi-lingual 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.

Nissim, M., C. Matheson, and J. Reid. 2004. Recognizing Geographical Entities in Scottish Historical Documents. In *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004.*

Papageorgiou, H., P. Prokopidis, V. Giouli, and S. Piperidis. 2000. A Unified POS Tagging Architecture and its Application to Greek. In *Proceedings of the 2nd Language and Resources Evaluation Conference.* 1455-1462. Athens, Greece.

Papageorgiou, H., P. Prokopidis, V. Giouli, I. Demiros, A. Konstantinidis, and S. Piperidis. 2002. Multi-level XML-based Corpus Annotation. In *Proceedings of the 3nd Language and Resources Evaluation Conference*.