Voula Giouli

Annotating corpora from various sources in the humanities domain: shortcomings and issues

In this paper, we present work aimed at the linguistic annotation of Greek corpora that belong to the humanities domain, the focus being on the methodological principles as well as the implementation framework adopted. This framework builds on an existing XML annotation platform that was initially developed in an Information Extraction setting and in order to cope with texts that pertain to domains such as news, administrative, economics, etc.; we elaborate on the initial steps taken towards customization of the tools.

1 Introduction

Over the last years, there has been a significant effort in creating various annotated corpora that have been made available in order to serve as training and evaluation benchmarks for Natural Language Processing (NLP) tasks even for the so-called "less-resourced" languages. These corpora are meant to model language used in specific domain-oriented applications. In this paper we present work aimed at the development and annotation of text corpora pertaining to disciplines in the humanities. Annotations have been carried out with the use of existing generic NLP tools that are currently customized so as to handle older and/or dialectical language varieties depicted in the corpus.

The paper is organized as follows. In section 2 we provide an overview of the corpus collection, regarding composition, size, features and the metadata schema employed for the representation of the digital content. In the following section, we present the levels of annotation that have been implemented so far, along with a suite of corresponding generic NLP modules for the Greek language, which have been used to initiate the annotation process. In section 4, after commenting on problems arising from the language varieties at hand, we present the steps taken so far for the customization of the afore-mentioned generic tools. The multilevel annotation tool has been employed for the extensive manual annotation of the data is presented in section 5, whereas initial remarks are discussed in section 6. Finally, conclusions are outlined in section 7.

2 Corpus description

The corpus hereby presented was initially developed in order to be integrated into a platform aimed to promote and highlight the cultural heritage of the Northern areas of Greece the focus being on literature and folklore. The intended infrastructure (corpora and platform) were targeted to a rather diverse audience ranging from students to teachers and the general public alike. Corpus collection followed a two-steps procedure thus reflecting the modular approach taken within the project it was initially intended for. At the first stage, texts were selected adhering to the genres of (a) literature, (b) folktales and legends, and (c) folklore texts (i.e., those commenting on and/or depicting a wide range of aspects of everyday human

activity such as traditions, customs, practices, spiritual beliefs in past eras). A set of pre-defined criteria conforming to specifications in the axis of place and time guided the design and content of the intended textual resource.

However, literature cannot be set apart from its era and the historical settings. Moreover, interpretations made by specialists (i.e., literary critics) are always sought for. To make therefore the platform as complete as possible, and a useful tool to prospect users, the initial collection was further coupled with texts that were intended to serve as accompanying material to the literary texts, namely authors' biographies, commentaries and literary criticism; on top of that, historical texts depicting the background of literary texts were also added to the collection. Within the intended project, the afore-mentioned criticism and historical texts had a two-fold purpose: (a) to be used as accompanying material, and (b) to guide the extraction of indexing terms, the ultimate goal being the development of a thesaurus that will enhance access to and retrieval of the primary data.

The so-collected corpus amounts to 304K words, and it covers a time span from the 19th century till the present day. Moreover, it represents a range of language varieties in the axes of time (i.e., contemporary, non-contemporary language) and place. More precisely, texts dated prior to 1976 (when *Dimotiki* was declared the official language of Greece) depict the non-contemporary language variety of "katharevousa", whereas, a number of literary and folklore texts depict the language variety spoken in the northern areas of Greece (*northern* dialect). Corpus composition and language coverage are depicted in Table 1 below:

	contemporary	***non-contemporary***	***dialectical***	***total***
literature	54K	57K	49K	160K
folktales	18K	-	23	41K
folklore	19K	22K	-	41K
historical	-	62K	-	62K
total	91K	141K	72K	304K

Table 1: Corpus composition

2.1 The metadata schema

To ensure easy access and re-usability of the corpus, a metadata scheme compliant with state-of-the-art standards was adopted, with certain modifications that cater for the peculiarities of the texts. The encoding scheme is compliant with the specifications of the Text Encoding Initiative (*http://www.tei-c.org*) (TEI Guidelines for Electronic Text Encoding and Interchange). To this end, metadata elements have been deployed reflecting bibliographical information that is primarily important for text identification with respect to text title, author, publisher, publication date, etc. (bibliographical information). Additionally, information on certain characteristics of the texts, such as language variety or sublanguage (contemporary/non-contemporary/idiomatic) was also added to the metadata descriptions manually.

In order to ensure documentation completeness and facilitate the inter-relation among primary data (i.e., literary texts) and the accompanying material (biographies, commentaries, criticism, etc), the documentation scheme has been extended accordingly so as to include such descriptive elements. Information regarding text type/genre and topic (where applicable) was also added manually on the grounds of generally accepted standards. To this end, *folklore* texts have been classified in accordance with the Classification scheme developed and used by the Library of Congress *(http://www.loc.gov/catdir/cpso/lcco/)*, whereas folktales categorization is conformant with the widely established Aarne-Thompson classification system (Aarne, 1961).

To keep track of the status and management of Intellectual Property Rights of the selected documents, appropriate metadata elements have been employed too.

From another perspective, the metadata scheme implemented in this project caters for the linguistic annotations that were provided for. The scheme employed 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*). This standard is compatible with TEI and can be mapped if considered appropriate. It has also been favored due to the fact that it is more appropriate for linguistically annotated corpora. On top of that, metadata elements inspired by the Dublin Core Metadata Initiative (DCMI) standard, including among others, *Annotator* (an entity responsible for providing the annotation content), *Subject* (what the annotation is about), *Resources* (the resources and tools that have been used in the annotation session), *Language* and *Date* (a date associated with the current session) are also included in the metadata headers.

Most of the aforementioned metadata descriptions were added manually to the texts and are kept separately from the primary data in an xml header that is to be deployed by the text management system for search and retrieval purposes. Moreover, metadata are stored separately from the raw data.

3 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.

However, corpus development within the NLP community is meaningless unless appropriate encodings or annotations are included that are designed to support different views of the language. To this end, to further enhance the textual collection, rendering it, thus, a useful resource to prospective users and researchers, further annotations at various levels of linguistic analysis were integrated into the primary textual material. These annotations served a two-fold purpose, that is, to enhance efficient indexing and retrieval of the textual documents, and to further facilitate the study of textual data and the elicitation of meaningful observations over the data.

3.1 Linguistic Annotation of texts

Linguistic annotation involves the following levels of analysis: (a) Part-of-Speech (POS) tagging and lemmatization, (b) surface syntactic analysis (chunking), (c) indexing with terms/keywords and Named Entities (NEs), (d) coreferece encoding, and (e) dependency annotation. These layers of linguistic annotations will be further elaborated in the remaining of this section. More precisely:

Part-of-Speech tagging (POS-tagging) is the first stage of linguistic analysis and involves the assignment of word class (part of speech) information coupled with more fine-grained morphosyntactic characteristics to every token in the text. Our scheme employs a PAROLE-compliant tagset. Surface syntactic analysis (chunking) consists in the recognition of non-recursive phrasal categories: adjectives, adverbs, prepositional phrases, nouns, verbs (chunks). Main as well as subordinate clauses were also recognized and labeled as appropriate.

Building on existing schemes developed for the annotation of NEs in texts, namely MUC-7 (Message Understanding Conference) and ACE (Automatic Content Extraction)), annotation at this level of linguistic analysis caters for the recognition and classification of the following types of entity names: person (PER), organization (ORG), location (LOC) and geopolitical entity (GPE). The generic schema also caters for the identification of numerical values: *(MONEY), (PERCENT),* and certain time expressions: *(DATE)* and *(TIME)* – yet, only the former was retained. Moreover, NE's of the type *(LOC)* were also assigned a subtype value, namely: geographical region (GEO) and facility (FAC). Though compatible in form with ACE, in that it retains most of the types and subtypes provided for by ACE, our classification schema differs in that disambiguation between *(LOC)* and *(GPE)* uses of names is being attempted.

At the next stage, terms were spotted and recognized. Conceived as the linguistic representation of concepts pertaining in a certain subject field, and being characterized by special reference "as opposed to words that function in general reference over a variety of codes" (Sager, 1980), terms and their identification were deemed meaningful only for the more "technical" texts in the collection, namely those pertaining to the domain folklore. Annotation at this level consisted in the selection of the word or word group that form a *simple-word* or *multi-word term in the given domain,* and their association with a pre-defined hierarchical list of topics. This list was created on the basis of the Library of Congress Classification scheme *(http://www.loc.gov/catdir/cpso/lcco/)*, and augmented on an as needed basis (see Fig. 1 below):

At the level of dependency annotation, the head-dependent relations among syntactic constituents were encoded, for representing the syntactic structure of a sentence. Grammatical roles that are identified and annotated include subjects, predicative complements, direct and indirect objects, prepositional phrases functioning as arguments or modifiers, and clausal arguments. Guidelines for the annotation of Modern Greek (MG) (Prokopidis, et al. 2005) and ancient Greek (Bamman et al. 2008) were taken into account. The latter were of particular importance for the annotation of texts written in the older language variety (*katharevousa*).

From the broad set of referential phenomena that characterize Greek language, we have focused on NP co-reference. In our work, two forms of co-reference have been accounted for: intra-sentential, in which case the co-referring expressions occur in the same sentence,

and inter-sentential, where a nominal expression refers to an entity mentioned in a previous sentence. The annotation involves: (a) the identification of *markables* in a sentence, that is, definite, indefinite and bare NPs, (b) the assignment of values to a set of attributes corresponding to their form (definite/indefinite/bare) and function (apposition, argument, etc), and (c) the identification of their antecedent. The interlinking of mentions of the same entity in a text results in the creation of *co-referential chains*. The anaphoric relation treated is that of identity. Our encoding scheme builds on the guidelines provided by the MATE project with certain modifications so as to cater for the particularities of the Greek language, as for example the fact that Greek is a pro-drop language.

3.2 Natural Language Processing tools

Annotations at almost all levels of linguistic analysis were performed semi-automatically (except for the last one that was applied manually), using a NLP pipeline developed at the Institute for Language and Speech Processing. The tools have been trained on Greek textual data from various sources (newspapers, internet, etc.) that cover domains such as finance, politics, sports, travel, etc. The main modules of this pipeline include a *tokenizer*, a *POS tagger and lemmatizer*, together with tools that recognize NEs and non-recursive syntactic units.

More precisely, 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 for the recognition of word and sentence boundaries, abbreviations, digits, and simple dates. To accomplish this task, we used the POS-tagger developed in-house (Papageorgiou et al. 2000) that is based on Brill's Transformation Based Learning architecture (Brill, 1997). Following POS tagging, lemmas retrieved from a Greek morphological lexicon were assigned to every word form.

A maximum-entropy Named Entity recognizer (Giouli et al. 2006) trained on financial and travel data identifies NEs of the afore-mentioned types (cf. above).

A term detection module (Georgantopoulos et al 2000) was then employed to identify terms in Greek text. It is a hybrid system comprising a regular expression-based term pattern grammar, and a statistical filter, used for the removal of terms lacking statistical evidence. Term Extractor functions in three pipelined stages: (a) POS 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.

In parallel, a module responsible for the automatic identification of grammatical relations has been employed that works on the basis of a pattern matching mechanism. The main resource used at this stage is a sub-categorization frames lexicon. The entries have been retrieved from a database containing sub-categorization information for the 5927 most frequent verbs, 4950 most frequent nouns, and 375 most frequent adjectives of a general purpose corpus.

4 Validation of the automatic processing

As it has already been mentioned, annotation was performed in most cases semi-automatically, that is, automatic processing using the afore-mentioned NLP tools followed by human validation. To minimize the effect of error transferring from previous levels to consecutive ones, the output of each processing component was manually validated prior to being fed to the next processing module. Moreover, due to the fact that the POS-tagger has been reported to achieve high accuracy levels (F-score 0.97) on standard texts, manual annotation was performed by two expert linguists only on the sub-corpus that deviated from the norm, that is, the non-contemporary and dialectical texts. Accuracy at the levels of NE and term annotation was even lower ranging from 0.21-0.63 depending on the text type. Moreover, the initial NE annotation schema was proved to be inadequate for the texts at hand.

For each annotation level, initial guidelines were provided to the linguists in charge of each annotation task. These guidelines were initially developed by expert linguists on the basis of existing encoding specifications in view of training generic NLP tools for a certain domain/text type. Within the current project, however, the initial specifications were appropriately revised so as to accommodate the peculiarities of the data at hand. For example, at the POS-tagging level, the dative case or the morphologically distinct subjunctive mood of the *katharevousa* (see below) should be accommodated for, the ultimate goal being the efficient description of the language variety used in the older texts. Similarly, NE annotation was meaningless in the current setting (literature, legends, and folktales) if it was intended for entities of the type (ORG). To this end, only NE's of the type PER and LOC were retained, and the specifications were modified so as to also include entities that are of interest in the texts at hand. The following new entity (sub-) types were defined, and our initial annotation scheme was revised accordingly:

- PER.human: Names of people, either dead or alive were further classified as human
- PER.animal: Names of animals fell into this subtype
- PER.fictional: Names of fictional characters were also tagged
- PER.other: All other animate entities that do not fall into the above subtypes were tagged as PER.other.

After a brief testing period of the new schemas/guidelines, and following the amendments or clarifications that were considered appropriate, samples by the annotators were collected and inter-annotator agreement was examined. Labels assigned by the two annotators were compared, and if the same label was assigned to the same spans of text by both annotators, it was counted as a match, otherwise not. By this measure, the average agreement score was counted around c. 90% for all levels of linguistic analysis.

As it has already been said, the major shortcoming in this procedure consisted in that the automatic processing of the textual data yielded very poor results especially in the cases of texts depicting language varieties that deviate from the norm. Manual annotation, on the other hand, aimed at re-training the tools is costly and time-consuming. To reduce manual effort, human validation has been performed on half of the data.

A close inspection over the data helped us to identify errors, and also to classify the sources of erroneous output so as to find appropriate solutions. A close inspection over the data has revealed the following as the main error-baring cases:

- **problematic/erroneous output from the Optical Character Recognition (OCR) module** resulting into various misspellings or even into a intelligible output;
- **encoding problems,** resulting from the conversion of initial documents to a format that is appropriate for the processing tools;
- **various misspellings or variant spellings;**
- **non-standard orthography and spelling variation** due to the language variety depicted in the literary works and/or the non-contemporary and idiomatic texts in the folklore domain.
- **word-formation and or declension** in accordance with the paradigm of the older/regional language variety.

4.1 Annotating non-contemporary and idiomatic words

The texts collected do not depict or represent a uniform variety of the Greek language. Instead, depending on the text type and the date of publication three main varieties are depicted: (a) *Modern Greek,* the official language of Greece, (b) *katharevousa*, and (c) a language exhibiting features of the *Northern Greek dialects.*

The situation of *diglossia*, i.e. the simultaneous existence of a vernacular and a high variety of the Greek language, was prominent from the birth of the new country until practically the end of the 20th century. Shortly after Greece was declared independent in 1830, the language issue was raised. The traditionalist, influenced by the Enlightenment ideal for a national language "argued for the resurrection of the classical Greek, uncontaminated by 'impure' admixtures with which it had been 'polluted' during its contacts" (Dendrinos, 2007). Their opponents, on the other hand, favored over the usage of the language actually spoken by the people. In between the two options, a third one advocated the use of the current language, 'purified' through its infusion with classical Greek in terms of morphology, syntax and vocabulary. The latter, which bore also the symbolic charge of continuation of Ancient Greek, prevailed, leading to this situation of diglossia. The high variety, *Katharevousa* (from *katharo,* meaning "clean"), an imitation of classical Greek was used in administration, education, science while the low variety, *Dimotiki,* was used in everyday informal communication, literature (although not by all authors) and primary education. This situation is reflected in those texts in the collection which were dated prior to 1976, and the prevalence of the (mainly the historical ones, most of the folklore texts and a few literary ones)

Literary and folklore texts, on the other hand, depict a language variety exhibiting all the characteristics that are present in the *Northern Greek dialects* (roughly covering the areas of central Greece, Thessaly, Macedonia, Epirus, Thrace, Euboea, and some islands in the Ionian and NE Aegean). The peculiarities of this language variety of Greek consist in deviations from the norm with respect to: (a) phonological features (the characteristic process of high-vowel (i, u) deletion in unstressed syllables leading to the creation of various consonant clusters), (b) syntactical features (the use of 'object' pronoun forms as indirect objects), and (c) morphological features.

All texts in the collection are appropriately encoded with respect to the language variety used. However, to keep track of the lexical entries deviating from the norm, words/word forms were appropriately tagged with this respect. This was especially important for texts which depict more than one language varieties shifting one language variety to another (norm, "*katharavousa*", and the *regional variety* with all possible combinations).

4.1 Towards resource customization

Annotations applied to the texts automatically were 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 expansion (in compliance to the PAROLE specifications) so as to capture the morpho-syntactic characteristics of the "katharevousa". To this end, the extended tagset caters characteristics such as the dative case in nouns, adjectives, articles, pronouns and in all the three genders, as well as the existence of participles, infinitives, and of the morphologically distinct subjunctive mood for verbs. This has resulted to the increase in the numbers of the allowed tags from 584 initially used for Modern Greek to 625 for the older language variety.
- enrichment of the morphological lexicon employed by the POS-tagger and/or revision in two axes: (a) inclusion of 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.
- revision of the annotation specifications at POS-tagging level so as to capture the peculiarities of the language variety at hand.
- Revision of the specifications set by MUC/ACE for the identification of NEs that are more relevant to the text types (cf. above). Additionally, new trigger words were manually selected for inclusion in the relevant tool.

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

5 A graphical user interface: Marker

An important component in the whole process of annotation was the usage of a flexible annotation environment called *Marker*. Marker is a Graphical User Interface that allows annotators to have simultaneous views of all levels of previous annotations, while working at a particular task. It supports annotations at the following levels of linguistic analysis: (a) morpho-syntax; (b) chunk and recursive phrases; (c) Named Entities; (c) term spotting and annotation; (d) coreference annotation, and (e) annotation of grammatical relations.

Classes of XML annotations that share a common vocabulary and structure (morphology, syntax, etc.) are described in DTD's. The Marker looks for the relevant DTD when initiating

an annotation session and configures the GUI appropriately by providing the needed functionality to the annotator. This dynamic process of building and customising a GUI on the fly (based on external DTD files) is currently restricted to simple elementary structures which however fulfill most of our current annotation needs. Additionally, a validation step is being performed ensuring that a particular instance is compliant with the pre-specified constraints in the DTD's. This environment also encompasses an editor which was extensively used for the editing/modification of the initial metadata and/or the rearrangement of their hierarchy in the schema. New annotation schemas were also implemented using the functionalities provided by the tool.

Within the current project, the tool has also served as an aid in our lexicographic work. Although it is not a proper lexicographic environment, it allows, at the level of term annotation for the inclusion of other information, such as definition, and reliability. The former was automatically retrieved along with lemma information and domain type/subtype facilitating, thus the population of the glossaries that were developed for the specific collection and domains covered.

6 Discussion

The textual collection that has been described above was primarily intended for laymen. As it has already been pointed out, the ultimate goal of the whole project was to create a set of language resources along with an infrastructure targeted to a wide and rather diverse audience. The application was aimed to serve as a teaching aid either in the domain of literature and folklore, or even in language teaching and learning. However, we argue that it can also be perceived as a pilot work that may guide future large-scale endeavors aimed equally at researchers as well. In this respect, 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 capabilities offered by NLP. The intended tools would be useful for a number of applications ranging from automatic indexing and retrieval of documents in specialized digital libraries, to the extraction of glossaries and the (comparative) study of word usage across writers, local communities, etc. to mention but a few.

However, the tendency for creating textual collections coupled with metadata for a variety of languages and language varieties, and the resulting need for portability and customization of generic NLP tools, has brought about the issue of a basic research infrastructure that goes beyond the needs of customary language technology (LT) applications. This need guides us to the notion of the Basic Language Resource Kit (BLARK), which refers to a core set of language resources and LT tools that are deemed essential not only to basic research in LT but also to the development of a variety of applications for a particular language (i.e., linguistically annotated text corpora, lexical resources, tools for linguistic annotation of tools, etc). And although this notion usually refers to modern standard languages and state-of-the-art applications, researchers are know starting to argue in favor of the idea of a BLARK that goes beyond the standard or modern usage of language, extending itself to one or more of the following axes of language

variation: (a) community (languages, dialects, sociolects), (b) subject, purpose or medium (topics, genres), (c) time (historical language stages) (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 conceived from the beginning as a contribution to a BLARK for the Greek language extended in the axes of time and community, the focus being at present on the creation of annotated corpora, the elaboration of annotation schemes, and the development/modification of accompanying lexical resources.

7 Conclusions

We have hereby presented work aimed at the annotation of specialized corpora comprising texts from the humanities disciplines. We have described the methodology adopted and the tools used, elaborating on the annotation schemas and the initial steps towards tool customization. Manual validation of the output of the automatic processing was intended for training the respective tools so as to handle texts in the domains and language varieties at hand.

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.

ACE. http://www.itl.nist.gov/iaui/894.01/tests/ace/

BAMMAN, D., AND CRANE, G. (2008). Guidelines for the Syntactic Annotation of the Ancient Greek Dependency Treebank (1.1). The Perseus Project, Tufts University. September 1, 2008.

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 Proc. of the 6th European Conference on Research and Advanced Technology for Digital Libraries. 613–625.

BORIN, L., AND M. FORSBERG (2008b). Something old, something new: A computationalmorphological description of Old Swedish. In LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008). 9-16. Marrakech: ELRA.

BORIN, L., D. KOKKINAKIS, AND L. J. OLSSON (2007). Naming the past: Named entity and animacy recognition in the 19th century Swedish literature. In *Proc. 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 *Proc. of LREC 2010.*

BROWNING, R. (1991). Medieval and Modern Greek. 1st edition 1962, 2nd edition 1983; Greek edition 1991 Athens: Papadima Publications.

CHRISTIDIS, A.F., ed. (2000). La Langue Grecque et ses Dialectes. Thessaloniki: Centre de la Langue Grecque.

CRANE, G. (2002). Cultural Heritage Digital Libraries: Needs and Components. In Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science. Vol. 2458. 51-60.

DAVIES, S., POESIO, M., BRUNESEAUX, F., ROMARY, L. (1998). Annotating coreference in dialogues: proposal for a scheme for MATE. First draft. Available at http://www.hcrc.ed.ac.uk/~poesio/MATE/anno_manual.html

DENDRINOS B., THEODOROPOULOU, M. (2007). Language issues and language policies in Greece. EFNIL, Riga http://www.efnil.org/documents/conference-publications/riga-2007/Riga-06-Dendrinos-Mother.pdf

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

GEORGANTOPOULOS, B, PIPERIDIS, S. 2000. *Term-based Identification of Sentences for Text Summarization*. In *Proceedings* of LREC 2000

GIOULI,V., KONSTANDINIDIS, A., DESYPRI, E., PAPAGEORGIOU, H. 2006. Multi-domain Multi-lingual Named Entity Recognition: Revisiting & Grounding the resources issue. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation, Geneva, Italy.*

NISSIM, M., C. MATHESON, AND J. REID (2004). Recognizing Geographical Entities in Scottish Historical Documents. In *Proc. of the Workshop on Geographic Information Retrieval at SIGIR 2004.*

PAPAGEORGIOU, H., PROKOPIDIS, P., GIOULI, V., PIPERIDIS, S. *2000*. A unified tagging Architecture and its Application to Greek. *In: Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece.*

PROKOPIDIS, P., DESIPRI, E., KOUTSOMBOGERA, M., PAPAGEORGIOU, H., PIPERIDIS, S. (2005). Theoretical and Practical Issues in the Construction of a Greek Dependency Treebank. In *Proceedings of The Fourth Workshop on Treebanks and Linguistic Theories* (TLT 2005). Barcelona, Spain.

RAYSON, P., D. ARCHER, A. BARON, AND N. SMITH (2007). Tagging historical corpora – the problem of spelling variation. In *Proc. of Digital Historical Corpora, Dagstuhl-Seminar 06491*. 3-8. International Conference and Research Center for Computer Science, Schloss, Dagstuhl, Wadern, Germany.

RAYSON, P., D. ARCHER, A. BARON, J. CULPEPER, AND N. SMITH (2007). Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proc. of Corpus Linguistics* 2007. Birmingham: University of Birmingham.

SAGER, J.C., DUNGWORTH, D., MCDONALD P. F. (1980). English Special Languages. Oscar Brandstetter Verlag KG - Wiesbaden.