(Re)thinking the BLARK for Ancient Greek

Federico Boschetti, Riccardo Del Gratta, Francesca Frontini, Fahad Khan, Monica Monachini

Institute for Computational Linguistics "A. Zampolli", CNR
Via Moruzzi, 1 - 56124 Pisa, ITALY
{name.surname}@ilc.cnr.it

Abstract

The aim of this paper is to discuss the Basic LAnguage Resource Kit (BLARK) for Ancient Greek, measuring the BLARK matrix against what is actually available for this language, and assessing its applicability to ancient languages in general. In addition, the BLARK and the FLaReNet recommendations are used to define priorities in the sector in close collaboration between philologists and the broader LRT community.

1. Introduction

The study of Ancient Greek and Latin is mainly distributed between Europe and North America, even if interest is increasing in other continents. A recent survey[1] provides an indication of the size of the potential community. We can estimate that those who study or have studied ancient Greek currently stand at more than one million and that for Latin there are at least five times more, since the results of the aforementioned survey are based only on secondary school data, excluding college and university students and scholars and data related to Greece.

The demographic make up of this linguistic community is very peculiar: it has no national identity, no native speakers (i.e. Ancient Greek and Latin are studied only as L2), and only rare new textual production. However, they have great importance for philological studies, linguistic analyses, and multilingual commentaries.

In this paper we focus on Ancient Greek, because Latin is a language that is leaving behind the less resources status, see (Passarotti, 2010).

2. Preliminaries

First of all, before we discuss a BLARK for Ancient Greek, as defined by (Binnenpoorte et al., 2002) and (Krauwer, 2003), we should focus our attention on some lower level issues: character encoding, font subsets, keyboard layouts.

2.1. Character Encoding

The Greek orthographic reform, which dates from 1982, reduced the high variety of diacritics (accents, breathing marks, subscript iota, diaeresis, possibly combined) in modern Greek to a monotonic accent and diaeresis.[2] Due to the limited expressivity of 8-bit encodings, modern (monotonic) Greek and ancient (polytonic) Greek have been treated in different ways by different communities. The former community is represented by Greek speakers that currently numbers ten million and it is attractive for commercial reasons; the latter is constituted by classicists (spread worldwide) that number less than one million and which is only academically relevant.

Even after the advent of Unicode, which supports both monotonic and polytonic orthographies, limitations and misunderstandings still exist. As illustrated in Table **??**, the monotonic accent and the polytonic acute accent have two different code points in the Unicode representation, however in many cases the polytonic acute accent is encoded as the monotonic accent. Moreover, the majority of fonts represent the monotonic accent in the same way as the polytonic acute accent, but (more correctly) some fonts represent them in distinct ways: the monotonic as a small vertical sign and the polytonic as a small ascending sign. Due to this confusion, conversion and adaptation is necessary for text retrieval systems, spell-checkers, etc.

2.2. Keyboard Layouts and Font Subsets

The main operating systems and text editors provide a Greek keyboard layout in two flavors: monotonic and polytonic, but unfortunately, due to the issues illustrated above, the polytonic acute accent is encoded as the monotonic one. Even worse, some operating systems for mobile devices and e-book readers do not provide polytonic keyboard layouts or other polytonic input methods at all, highly compromising the usability of Ancient Greek resources and tools. Many Unicode fonts provide the Greek (i.e. monotonic) subset, but not the Greek Extended (i.e. polytonic) subset, jeopardizing the correct visualization of Ancient Greek texts. Some operative systems for mobile devices have no pre-installed fonts with the Greek Extended subset.

Both academic and commercial stakeholders should be aware that polytonic Greek is necessary not only for classical studies, but also to digitize and process modern Greek texts produced before the orthographic reform: there are millions of documents written in polytonic Greek from antiquity up to 1982, including newspapers, commercial agreements, out of copyright best sellers.

3. Data, Technologies and Applications for Ancient Greek

Table 1, based on the scheme created by (Binnenpoorte et al., 2002) – and adapted to Arabic by (Maegaard et al.,

[1] According to the survey provided by Emily Franzini (Open Philology Project) at <http://www.dh.uni-leipzig.de/wo/update-total-number-of-secondary-level-students-studying-latin-and-ancient-greek-in-the-world>, students in secondary school that study Latin in the countries for which data have been gathered, are 3,579,767, whereas students of Ancient Greek are 736,278.

[2] i.e. from polytonic εὐχαριστῶ to monotonic ευχαριστώ

2006) –, illustrates a hypothesis to evaluate, respectively, the importance of Ancient Greek digital Resources for software modules, and the importance of the modules for the applications. In the current scenario, modules can be software components implemented through a specific programming language (e.g. Java), or web-services, which remotely provide an atomic functionality. Applications can be desktop and web application, or mobile apps.

The next subsections illustrate resources, modules and applications for the study of Ancient Greek relevant for the BLARK, in order to create better conditions for research, education and development in language and speech technology. We decide to mention open data and open source tools only. Further information can be found in (Babeu, 2011).

3.1. Data

Training sets for OCR. OCR represents the real bottleneck for Ancient Greek digital resources. In the last decade training sets for the OCR engines *tesseract*[3] and *gamera*[4] have been collected, in order to enhance the accuracy of the recognition. Further improvements are necessary in the layout analysis, in particular for critical editions with a complex layout (original text, modern translation, critical apparatus and possibly exegetical notes in separate text boxes per page).

Textual Corpora. The largest unannotated and annotated (in TEI[5] P4 or P5) text collections under an open license are provided by the Perseus Project.[6] The Perseus Project Digital Library is constantly updated with new texts acquired by OCR and manually corrected, because even digital texts currently available under copyright must be independently acquired from printed editions, in order to provide the community with texts not only searchable, but fully actionable under open licenses.

Variants and Multiple Editions. Ancient Greek Texts are attested in ancient and medieval witnesses (papyri, manuscripts, etc.) or reconstructed by scholars in modern editions through collation of the primary sources and conjectures for the emendation of probable errors. Variant readings observed in the manuscripts and conjectures contained in modern editions can be mapped on the reference edition or compared with concurrent variants and conjectures. Many out of copyright critical editions are available at the Internet Archive website,[7] but they need to be processed by OCR. The quality of the original printed edition and the quality of digital images affect the accuracy of the acquisition.

Multimedia Primary Sources. The images of primary sources (epigraphs, papyri, manuscripts, but also printed editions etc.) are important in order to compare the transcription of the text and the different interpretations in case of palaeographic uncertainty. One of the highest quality projects on primary sources and transcriptions is the Multitext Homer Project.[8] A BLARK for ancient languages should take into account the linkage between text and its material support: archiving and exploitation of ancient artifacts should play a key role in the basic resource kit.

Multilingual Secondary Literature. Secondary literature, such as commentaries, articles and monographs, are necessary for philological activity, because each suggestion for emendation and each interpretation of the text must be compared with the critical activities of previous scholars. Google Books[9] and Internet Archive are mines of relevant, out of copyright, documents. Even in this case the acquisition by OCR is necessary, although the recognition of Latin characters is less challenging than the recognition of polytonic Greek.

Lexica, Thesauri, Ontologies and WordNets. The Perseus Project provides the Liddell Scott Jones lexicon and other bilingual dictionaries, which are available also through *Logeion*[10]. The digital Liddell Scott Jones dictionary is fully annotated according to the TEI[11] guidelines and it is one of the best lexicographical resources available online for the antiquity. At CNR-ILC we are developing the Ancient Greek WordNet (AGWN)[12], as described in (Bizzoni et al., 2014), in collaboration with the Alpheios Project[13].

Parallel Multilingual Corpora. Parallel Multilingual Corpora are a valuable source of linguistic knowledge which forms the basis of NLP techniques. They have been used to create the Dynamic Lexicon[14], a set of automatically built bilingual dictionaries (Greek/English and Latin/English), obtained by making use of parallel texts - source texts in Greek or Latin aligned with their English translations - along with the syntactic data encoded in treebanks. Parallel corpora are proven to be useful for didactic purposes, as well as for the study of scholarly interpretations and their diachronic variations.

3.2. Technologies

Modules for Ancient Greek treatment should be focused on the following four areas.

Acquisition. After the pioneering phase of text digitization by typewriting, as explained in (Boschetti et al., 2009), the OCR technology is ripe to be applied to Ancient Greek.[15] Current efforts are focused on the acquisition of low quality page images and critical apparatus from printed scholarly editions. Clearly, spell-checking and hyphenation recognition are crucial for the accuracy improvement of OCR technologies, as well as transversal features needed by didactic and academic applications.

[3] Tesseract is available at <https://github.com/tesseract-ocr>; Ancient Greek OCR training sets are available at <https://github.com/tesseract-ocr/langdata>.

[4] Gamera is available at <http://gamera.informatik.hsnr.de/addons/ocr4gamera>; Ancient Greek training sets for Gamera are available through <http://heml.mta.ca/lace>

[5] <http://www.tei-c.org>

[6] <http://perseus.tufts.edu>

[7] <http://www.archive.org>

[8] <http://www.homermultitext.org>

[9] <http://books.google.com>

[10] <http://logeion.uchicago.edu>

[11] <http://www.tei-c.org>

[12] The AGWN has been automatically created and its validation is still undergoing

[13] <http://alpheios.net>.

[14] <http://nlp.perseus.tufts.edu/lexicon>

[15] See Lace Project: <http://heml.mta.ca/lace>

Linguistic Analyses. The acquired texts should be linguistically analyzed at least at the following levels: phonemic, prosodic and metrical, morphosyntactic, semantic and referential.

Phonemic and prosodic analyses, as well as the metrical analysis for poetical texts (Pavese and Boschetti, 2003), are relevant for the stylistic study of literary texts, in order to identify the peculiarities of specific authors. At least for non lyrical verses, the automated metrical analysis is highly accurate.

Lemmatization and morphological analysis are necessary for the majority of the higher level linguistic modules, such as syntactic parsing, word sense disambiguation etc. Ancient Greek texts can be processed by *Morpheus* (Crane, 1991) with satisfactory accuracy, even if POS tagging could further improve it. In the future, it would be important to go beyond lemmatization, by the identification of single morphemes in derivative and compound words, in order to group etymological families of words.

According to the typical linguistic analysis chain, on top of lemmatization we should add the results of automated syntactic analysis, which however is still very challenging. The results of the first experiments have been provided by (**?**), which make use of the Ancient Greek Dependency Treebank as a training set. Word Sense Disambiguation could be facilitated by a combination of techniques that exploit distributional semantics, for contextual information, and the Ancient Greek WordNet (Bizzoni et al., 2014), for the identification of semantic relations. Finally, multilingual alignment is useful to extract relevant word-to-word, word-to-phrase and phrase-to-phrase translations from textual corpora. This is a complementary, syntagmatic approach to the paradigmatic approach of pair extraction from bilingual dictionaries.

Scholarly Editing. In addition to standard linguistic processing, textual corpora of historical languages need at least two specific philological modules. The former is necessary to manage variant readings attested in different witnesses (e.g. papyri, manuscripts, etc.), in order to visualize multiple documents and to compare different versions of the same text. The latter manages the secondary literature (e.g. commentaries, journals, etc.) connected to the primary sources, by extracting canonical citations, as explained in (Romanello, 2013).

Speech Recognition and Text-to-Speech. Our survey ends with few comments on speech acquisition and production. Speech recognition for Ancient Greek, to the best of our knowledge, is a neglected field, even if it is useful for multimodal OCR correction. Text-to-speech to improve accessibility of classical studies to blind students and scholars are under development[16]. Metrical and non metrical text-to-speech can be exploited for didactic reasons.

3.3. Applications

Education and Acquisition. As stated in Section 1., the impact of classical languages in western education is not homogeneous, but concentrates in some countries where it is very relevant. For instance, the Perseus and the Open Philology[17] Projects use the Computer Assisted Language Learning (CALL) systems to provide students with software to learn Ancient Greek.[18] Bilingual corpora constituted by original Greek texts and their translations, not only are useful for educational purposes, but they are also crucial for the improvement of lexico-semantic resources.

Distant and close reading. Distant reading (Moretti, 2013), identifies the scholarly activity addressed to the literary works as a whole, in order to explore, by statistical techniques and visualization tools, clusters of texts, interesting outliers, etc.; while close reading identifies scholarly activities focused on narrow portions of text to analyze and interpret. If the former can be addressed using pure (statistical) linguistic methods, the latter needs all the data and all the modules available.

Philological activities The digital scholarly edition is a dynamic view of diplomatic and critical editions, where the former are an accurate transcription and description of textual phenomena in the primary sources and the latter are the result of primary source collations and conjectural emendations, with variants recorded in the critical apparatus. Finally scholarly Information Retrieval extends text retrieval to variants and multiple editions and exploits any level of annotation, in order to filter data to provide results relevant for philological investigations.

4. Defining priorities for the field

BLARK is part of a more general set of recommendations for the development and progress of LRTs in Europe, issued by the FLaReNet Project (Soria et al., 2014). Despite the differences between modern and classical languages, many of these recommendations can also be applicable to the latter.[19] The FLaReNet strategic agenda is organised along various dimensions; we highlight here the aspects that are particularly relevant for Ancient Greek.

The use of standards for annotation and representation of language data is the key to **interoperability** which allows one to 'go green', i.e. facilitating **sharing and reusing** of data as well as **repurposing** of existing technologies for Ancient Greek in close **collaboration** between philologists and the broader LRT community. For this purpose, existing standards and best practices in philology should be collected and clearly documented. In this sense, **documentation** is of the utmost importance; resources for Ancient Greek need to be regularly surveyed and described with the adequate **metadata** available in registers of the LRT community, e.g. the CLARIN-VLO, the METASHARE platform, etc., thus amplifying **visibility**. Moreover the proper **citability** of resources for Ancient Greek should also be a main concern, and scholars should be aware of the current discussion around a standard citation framework for LRT in general[20]. This would help

[16] Our current source is a private communication.

[17] <http://www.dh.uni-leipzig.de/wo/open-philology-project>

[18] Through CALL students also enlarge the treebanks with morphosyntactic annotations.

[19] Similar approaches have been already attempted for minority language (Soria et al., 2013).

[20] Cf. the ELRA ISLRN, www.islrn.org. Some initiatives to monitor resource usage are also already being adopted, such as the LRE Map (Calzolari et al., 2010), that might be adopted in

them to ensure **recognition** and measure the research **impact** of their resources. As for **sustainability**, both philologists and funding agencies of the sector should be made aware of the importance of ensuring the data life-cycle of the resources they produce. Rather than adopting self-made solutions, philologists should align themselves with the guidelines in the production, preservation and access of LRT, in particular by relying on what resource infrastructures such as CLARIN or DARIAH may offer to them in term of standardized services, legal solutions, and long term preservation and access.

5. Final remarks

The BLARK for a historical language, such as Ancient Greek, crucially needs to take into account philological aspects, e.g. variant management, multiple editions, secondary literature, etc., usually neglected by computational linguists focused on modern languages. In this contribution we tried to identify on one hand the low level bottlenecks that prevent the creation and the exploitation of textual and linguistic resources for Ancient Greek, and on the other hand we tried to list resources, modules and applications peculiar to historical languages in a linguistic and philological perspective. The BLARK proposed in this paper needs to be validated by the community of scholars involved in collaborative and cooperative philology through assessment questionnaires. Finally we provided some FLaReNet-style recommendations for the promotion of LRT for Ancient Greek. In perspective, we envisage the possibility of extending this work into a White Paper for Ancient Languages, along the lines of the META-NET studies of Europe's languages in the Digital Age(Rehm et al., 2014).

Last but not least, we would like to thank Harry Diakoff (Alpheios Project) for his support to this paper.

6. References

Babeu, Alison, 2011. *"Rome Wasn't Digitized in a Day": Building a Cyberinfrastructure for Digital Classics*. Council on Library and Information Resources.

Binnenpoorte, D., F. De Friend, J. Sturm, W. Daelemans, and C. Cucchiari, 2002. A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch. In *Proceedings LREC 2002, (Third International Conference on Language Resources and Evaluation)*. Las Palmas de Gran Canaria, Spain.

Bizzoni, Yuri, Federico Boschetti, Riccardo Del Gratta, Harry Diakoff, Monica Monachini, and Gregory Crane, 2014. The Making of Ancient Greek WordNet. In *Proceedings of Language Resources and Evaluation Conference, Iceland*.

Boschetti, Federico, Matteo Romanello, Alison Babeu, David Bamman, and Gregory Crane, 2009. Improving OCR accuracy for classical critical editions. In *Research and Advanced Technology for Digital Libraries*. Springer, pages 156–167.

Calzolari, Nicoletta, Claudia Soria, Riccardo Del Gratta, Sara Goggi, Valeria Quochi, Itene Russo, Khalid Choukri, Joseph Mariani, and Stelios Piperidis, 2010. The LREC Map of Language Resources and Technologies. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010)*. Valletta, Malta.

Crane, Gregory, 1991. Generating and parsing classical Greek. *Literary and Linguistic Computing*, 6(4):243–245.

Krauwer, Steven, 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources. *Proceedings of SPECOM 2003*:8–15.

Maegaard, Bente, Steven Krauwer, Khalid Choukri, and L Jørgensen, 2006. The BLaRK concept and BLaRK for Arabic. In *Fifth International Conference on Language Resources and Evaluation, LREC'06*.

Moretti, Franco, 2013. *Distant reading*. Verso Books.

Passarotti, Marco, 2010. Leaving behind the less-resourced status. The case of Latin through the experience of the Index Thomisticus Treebank. *Proceedings of the 7th SaLTMiL Workshop on the creation and use of basic lexical resources for less-resourced languages, LREC 2010, La Valletta, Malta*.

Pavese, Carlo Odo and Federico Boschetti, 2003. *A Complete Formular Analysis of the Homeric Poems*. Amsterdam.

Rehm, Georg, Hans Uszkoreit, Ido Dagan, Vartkes Goetcherian, Mehmet Ugur Dogan, Coskun Mermer, TamÃ¡s VÃ¡radi, Sabine Kirchmeier-Andersen, Gerhard Stickel, Meirion Prys Jones, Stefan Oeter, and Sigve Gramstad, 2014. An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age". In *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014)*. Reykjavik, Iceland.

Romanello, Matteo, 2013. Creating an annotated corpus for extracting canonical citations from classics-related texts by using active annotation. In *Computational Linguistics and Intelligent Text Processing*. Springer, pages 60–76.

Soria, Claudia, Nicoletta Calzolari, Monica Monachini, Valeria Quochi, Núria Bel, Khalid Choukri, Joseph Mariani, Jan Odijk, and Stelios Piperidis, 2014. The language resource strategic agenda: the flarenet synthesis of community recommendations. *Language Resources and Evaluation*, 48(4):753–775.

Soria, Claudia, Joseph Mariani, and Carlo Zoli, 2013. Dwarfs sitting on the giants' shoulders – how LTs for regional and minority languages can benefit from piggybacking major languages. In M.J. Norris, E. Anonby, M-O. Junker, N. Ostler, and D. Patrick (eds.), *Proceedings of the XVII FEL Conference*. Ottawa, Canada. 1-4 October, Carleton University.

conferences of the sector.

MODULES	DATA													APPLICATIONS					
	Training sets for OCR	Unannotated Corpora	Annotated Corpora	*Unannotated Variants*	*Annotated Variants*	*Unannotated Multiple Editions*	*Annotated Multiple Editions*	*Multimedia Primary Sources*	*Multilingual Secondary Literature*	Monolingual Lexica	Multi/bilingual Lexica	Thesauri, ontologies, wordnets	Parallel Multilingual Corpora	CALL	*Digital Scholarly Edition*	*Scholarly Information Retrieval*	Translation	*Distant Reading*	*Close Reading*
Language Technology																			
OCR	++	+		+		+		++	+					+	++	+	+	+	+
Spell-checking and Hyphenation		++	+	++	++	+								++	++				
Phonemic Analysis		++	++											+				+	+
Prosodic Analysis		++	++							+	+			+				+	+
Metrical Analysis		++	++							+	+			++	+			+	++
Lemmatization		++	++	++	++	++	++			++	++	+		++	++	++	++	++	++
(Etymological) Morphemic Analysis		++	++	++	++	++	++			++	++	+		++		+		+	+
Morphological Analysis		++	++	++	++	++	++			++	++	+		++			+	+	++
POS Tagging		+	++											++	++	++	+	+	++
Syntactic Parsing		+	++							+	+	+	+	++		++	+	+	++
Named Entity Recognition		+	++							+	+	+	+	++		++	+	+	++
Term Extraction		+	++							+	+	+	+	++		+	+	+	++
Word Sense Disambiguation		+	+							+	++	++	+	++		+	+	+	++
Multilingual Alignment		++	++										++	++			++	+	++
Variant Management		++	++	++	++	++	++						++	++	++	+			++
Secondary Literature Management		++	++						++				++	++	++	+	+	+	++
Speech Technology																			
Speech Recognition			++											+	+				
Non-native Speech Recognition			++											+	+				
Prosody Recognition			++											+	+				
Text-to-Speech			++	+										++					
Metrical Text-to-Speech			++	+										++					

Table 1: Importance of data for modules and importance of modules for applications related to Ancient Greek (according to (Krauwer, 2003), + means relevant and ++ means important modules. Data and applications specific to philological and literary studies are highlighted in italics)