Abstract The papers in this issue make two complementary assertions: first, language and linguistic sources are a key element of human cultural heritage and, second, we need to integrate the ancient goals of philology with rapidly emerging methods from fields such as Corpus and Computational Linguistics. The first 15,000,000 volumes digitized by Google contained data from more than 400 languages covering more than four thousand years of the human record. We need to develop methods to explore linguistic changes and the ideas that languages encode as these evolve and circulate over millennia and on a global scale.

Categories and Subject Descriptors: XXX – to insert I 2.7 [Artificial Intelligence] – Natural language processing – Speech recognition and synthesis.

General Terms: XXX Computational Linguistics, Corpus Linguistics, Philology

Additional Key Words and Phrases: XXX

1. INTRODUCTION

This issue of the Journal of Computing and Cultural Heritage builds upon a workshop held at Tufts University in January 2010 and jointly funded by the US National Endowment for the Humanities and the German Science Foundation. The workshop brought together not only researchers from the United States and Germany but also from the fields of Corpus and Computational Linguistics and of Classical Philology. From the papers given at the workshop we chose three for a special issue of JOCCH on “Corpus and Computational Linguistics, Philology and the Linguistic Heritage of Humanity.”

The papers in this issue make two complementary assertions: first, language and linguistic sources are a key element of human cultural heritage and, second, we need to integrate the ancient goals of philology with rapidly emerging methods from fields such as Corpus and Computational Linguistics.
2. LANGUAGE IS CULTURAL HERITAGE

Language and culture are intimately connected and we have an obligation to explore the cultural heritage of humanity with every resource at our disposal. We may for pragmatic reasons confine our immediate work to 3D modeling of historical space or to the analysis of geographic data but in studying the cultural heritage of humanity we must explore every element of lived experience and exhaust every aspect of the human record, including both the material and the linguistic record.

This idea is hardly new\(^1\) but the pressure to focus and to specialize our work means that we must constantly remind ourselves that the study of the past must be broader and deeper than any single discipline can encompass. Two hundred years ago, in 1811, August Boeckh brought this idea with him when he took up a position at the Humboldt University in Berlin. Reacting to a narrow vision of linguistic work as the reconstruction and interpretation of our surviving sources, he asserted that, even insofar as he himself focused primarily upon language, his proper subject was the totius antiquitatis cognitio, the understanding of the past in its entirety (Boeckh 1965, 1968, Calder 1991). Language might be the focus of his expertise and his individual contributions, but the linguistic record was never more than one element from the lived life of past generations.

Boeckh was a philologist, a term that may seem quaint even to those researchers who recognize it in the early twenty-first century. As recently as 1956, the situation was different: at the center of the ground-breaking film, Forbidden Planet, an elite rescue mission is dispatched to learn the fate of the Bellerophon mission. When they arrive on a distant planet, they discover that Dr. Edward Morbius, who has created for himself and his daughter a distant Eden, including a robot that is, “beyond the combined resources of all Earth's physical science.” Dr. Morbius is not an engineer, a physicist, or biologist but a “philologist, an expert in words and languages, their origins and meanings.”\(^2\) When a decade after the second world-war, the producers of this film chose in the philologist a intellectual center for voracious and unbounded intellectual curiosity, we hear the fading echo from researchers such as Boeckh, who developed in Germany a model for the research university upon which Americans would build.

\(^1\) The first electronic corpus – the works of Saint Thomas Aquinus, digitized in the 1940s by Roberto Busa, an Italian monk and philosopher – already worked with this assumption (Busa 1974, 1980).

In a globalized world, communities with heterogeneous languages and cultures, previously quarantined by distance, now interact across space in real-time. There are no significant reserves of oil in Afghanistan and Israel is a long way off. Young men and women from the United States, Germany and elsewhere are in harm’s way as we write this introduction because of clashing perspectives on the past and its bearing upon the present. Terrorism constitutes, in some measure, a violent response to the fear that various pasts will vanish in a homogenized culture of technocratic elites and consumer subjects. Two centuries after Boeckh arrived in Berlin, the study of Greco-Roman antiquity in particular may no longer stand at the center of academic activity, but our need to understand the past and its relationship to the present has, if anything, grown more pressing.

Whether or not we choose, as did Boeckh and many others, to employ the term philology or not, the impulse behind Boeckh’s vision is extraordinarily important. Instead of the Greco-Roman world alone, we must think in terms of the cultural heritage of humanity as a whole. This does not exclude the Greco-Roman world but rather situates it within a network of cultures that have co-evolved over long periods of time and established resilient systems that shape the present and the future. Students of Saddam Hussein would have been well-served had they paid more attention to the continuities between his rule and that of strong men such as Hammurabi and Nebuchadnezzar thousands of years before, while it is unclear how well we can understand the evolving role of the Chinese communist party without being able to evaluate the continuities with imperial bureaucratic rule and Confucian culture.

Publications with titles such as “From Plato to Nato” do not consider that fact that Plato and Aristotle are as much the cultural heritage of the Muslim as of the Judaeo-Christian world and that the Islamic Republic of Iran, with its philosopher king, its council of guardians and its carefully constrained democratic institutions, reflects an Islamic interpretation of the critique that Plato leveled on democracy in his Republic. When American journalist Robin Wright first met Iranian President Khatami in 1998, he quoted Plato and asked “what is justice?” to a bemused group of reporters, who had “all come to talk about issues a bit more pressing than ancient Greek philosophy.” (Wright 2008, 304). He also called for a dialogue, rather than a clash, of civilizations – a call that the violent events of 9/11 preempted. But if that call was disrupted, the need for such a dialogue of civilizations has grown, if anything, stronger.

Violence is, however, not the only barrier to a dialogue of civilizations. Language remains the greatest obstacle to human communication. We can reconstruct historical spaces with astonishing verisimilitude. Less than twenty-four hours of commercial air flights can place us on the other side of the globe and some governments could place forces anywhere on the earth. But acquiring the ability to understand and communicate directly in Chinese or Arabic remains, for the most part, as challenging now as it was a generation or a thousand years ago. UNESCO estimates that there are 6,700 languages spoken today, half of which are endangered. If we restrict ourselves to languages with a significant presence in printed literature, more than four hundred languages appear in the

---

first 15 million books digitized by Google. The official languages of India alone number 18,\textsuperscript{4} as well as four official Classical languages (Tamil, Sanskrit, Kannada, and Telugu).

If we are to evolve as a global society, we need to reorganize our approach to language. This has obvious implications for the educational structures in every nation but we must also confront the inherent limitations of human cognition. We simply cannot study, much less master, more than a tiny subset of human languages. All participants in modern networked society thus interact, directly or indirectly, with people who speak languages that they will never understand – imagine, as you read these words, the web of trade and production that lie behind the clothes that you wear, the food that you eat, the home in which you live, and the network on which this essay was produced.

This issue documents early steps in a new philology, one that applies the intellectual goals espoused by Boeckh to the full range of cultural heritage sources now becoming available to us and to four thousand years of the human linguistic record in particular. A field such as Classical Studies has an opportunity to redefine itself to include every lingua franca and every language that became a major instrument for the development and preservation of culture. The three great Classical languages of the Indo-European tradition, including as they have not only Greek and Latin but also Sanskrit, have always transcended the boundaries of modernity and of western culture. As other languages such as Classical Chinese and Arabic, Tamil, Kannada and Telugu from India, Tibetan, the cuneiform languages of Mesopotamia and the various forms of Egyptian from Hieroglyphics through Demotic and Coptic, begin to emerge, Greek and Latin cease to stand alone as the Classical languages but rather to assume a new – and far more significant -- role as major components in a global network of cultural systems.

The implications of such a shift are substantial for research and for education. We need to rethink the questions that we ask in light of the contributions that we wish to make to human intellectual life and each of these depends upon the technological infrastructure upon which we depend. This issue suggests a few of the changes now already visible.

3 CULTURAL HERITAGE AND CORPUS LINGUISTICS

Students of the past depend, by definition, upon fixed corpora and we must learn to draw upon aspects of corpus linguistics. At the same time, even if we confine ourselves to historical corpora much smaller than the flood of linguistic data pouring onto the World Wide Web, these historical corpora are often far larger than we can ever analyze with manual methods. We must also exploit every aspect of computational linguistics, modeling phenomena. The first three articles in this issue deal with computational, largely automatic analysis of large historical corpora. The final article describes a multi-layer architecture for deeply annotated historical corpora.

In the first article, David Bamman and David Smith describe how they extracted from the first million books made available for public download over the Web a collection of Latin that includes c. 1 billion words. Their discussion covers some of the technical challenges that they faced. About 20\% of the books listed as being in Latin were in fact in some other language (often Greek editions that had Latin title pages) while the use of a

\textsuperscript{4} Constitution of India. Articles 29, 30, 120, 210, 343-351 as amended in the 21st and 71st Amendments.

language model to scan the rest of the corpus showed that existing language metadata missed about 20% of those books that were in Latin. More importantly, Bamman and Smith address a limitation in the library metadata itself. The catalogue records described the publication date of the individual books rather than the period when the works themselves were produced – thus, the date for an 1887 edition of Cicero would situate it in the 19th century rather than the 1st century BCE. A team of research assistants went through c. 1/3 of the Latin books, creating a collection of c. 9,000 books for which reasonable production dates are available. The result is a serviceable diachronic corpus that includes 385 million words and covers more than 2000 years. While this corpus is only the starting point for much larger collections (Google has digitized more than ten times as many books as the 1.5 million sample from the Internet Archive), it already challenges students of this language to look beyond the 10 million or so words in the canon of Classical Latin and to study more than two millennia of Latin as a lingua franca at the heart of European culture.

In the second article David Mimno addresses the challenge of detecting ideas within and across languages. He takes as his starting point one hundred years of Classical scholarship as it appears in twenty-four journals in classical philology and archaeology available from JStor. Mimno uses statistical methods to detect co-occurring clusters of words that reflect coherent topics. Thus, the cluster “case language languages example nominative English genitives means” clearly captures a focus on grammar and language and, in turn, allows us to trace the shift away from grammatical topics in publications over a one hundred year period. By contrast, “woman women men man family young” emerges as a cluster that captures an interest in gender that explodes in the latter part of the twentieth century. While English is the primary language in these journals, enough of the publications were in German so that Mimno could detect different patterns in the two linguistic subcorpora. The corresponding gender topic in German (frauen männer gehört mädchen männlichen frau) increases but not to the same degree, while topics in English and German associated with ancient law are consistently more prominent in German than in English. The comparison between German and English scales to other languages – indeed, as collections grow larger, the statistical methods for the most part grow more effective. Mimno’s paper illustrates methods whereby we can ultimately trace ideas not only through a multilingual discipline such as Classical Studies but as they circulate across hundreds of languages and thousands of years.

The third article, by Hagen Hirschmann, Anke Lüdeling, and Amir Zeldes shows yet another aspect of working with historical corpora. It is not concerned with the automatic analysis of large corpora but with careful manual annotation of small, handcrafted corpora. In some sense this mirrors the ‘traditional’ philological way of working with historical data – in another sense it is completely different. Hirschmann and colleagues, using a small diachronic treebank of historical German to discuss a language change phenomenon (the development of the perfect), show how (qualitative) annotation and quantitative analysis go together. They show that only a sophisticated multi-layer standoff architecture that allows annotation in different formats on all linguistic levels is adequate when historical resources are to be used collaboratively and exploratively.

REFERENCES

BOECKH, A. 1966. Enzyklopädie und Methodenlehre der philologischen Wissenschaft, edited by Ernst Bratuscheck (Darmstadt, ed. 3) 25: “die Erkenntnis des Alterthums in seinem ganzen Umfange” =