# White Paper Report

Report ID: 104046

Application Number: HW5001009

Project Director: Gregory Crane (gregory.crane@tufts.edu)

Institution: Tufts University

Reporting Period: 5/1/2009-3/31/2013

Report Due: 6/30/2013

Date Submitted: 7/2/2013

Final Report and White Paper

HW-50010-09

# New methods for working with old languages: Corpus Linguistics and the future of Textual Scholarship

Gregory R. Crane

Tufts University

Submitted July 2, 2013

# DFG/NEH Bilateral Workshop

This report describes the project activities, the project outcomes and includes lessons-learned.

## Project Activities

All humanists now enjoy the prospect of working with very large collections of materials: Google, the Internet, the Hathi Trust and other organizations have assembled millions of digitized books available.  We have far more digital textual material available than was ever the case before. No one really knows how to exploit these rapidly expanding resources.

We conducted two workshops that brought scholars from Germany and the United States, from Classics, Corpus Linguistics and other areas of the humanities together to synthesize the state of the art to date, to develop immediate collaborations, and to publish concrete descriptions of what we can do now and where we can focus our research. Anke Lüdeling, professor of Corpus Linguistics at the Humboldt University, and Gregory Crane, Professor of Classics, Winnick Family Chair of Technology and Entrepreneurship, had already participated in joint seminars in Germany and the United States together before and ran a joint workshop on “What do you with a million books?” at Humboldt University in March 2008. Their on-going collaboration suggested to each the opportunities that scholars working with heavily researched corpora of historic languages and corpus linguists generally offer to one another. The DFG/NEH partnership has provided a unique opportunity to realize the benefits from such a collaboration.

The first workshop, held at Tufts University in January 2010,[1] focused on corpus linguistics and the field of classics and synthesized the automated methods now being applied in Germany and the United States that are of immediate relevance to historical languages such as Greek and Latin.  This workshop included classicists working at the University of Chicago, Harvard University, Harvard’s Center for Hellenic Studies in Washington, DC, Northwestern, and the host institution, Tufts University.  German participants represented projects with a focus on Greek and Latin and a commitment to automated methods such as eAqua, and Teuchos, as well as the Goettingen-based TextGrid project, which applies similar methods to German literature.

The second workshop, held at Humboldt University in Berlin in January 2011,[2] shifted the focus to the use of corpora for language instruction. The rise of linguistic corpora and of analytical methods from corpus linguistics has begun to open up new pathways for language learning. This workshop examined applications for students of both modern languages (such as English, German, Chinese and Arabic) and of historical languages (such as Greek and Latin) for which no native speakers survive.

[1] The agenda is available at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/events-en/nehdfg/

[2] http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/events-en/CTLL/CTLL_main/.

Students could, for example, define corpora that represent those areas of the language on which they choose to focus – these could be documents from news media or canonical texts. In this workshop, we discussed methods whereby students can assess their ability to acquire and then apply knowledge from these corpora as they work with new linguistic sources. We also investigated ways for faculty to assess the competence of students exploring disparate genres within a given language, fostering scalable assessment of students who pursue different pathways.

We also looked at the applications of annotations. How can learners benefit from developing and/or executing system annotation of linguistic corpora? How much can students learn about linguistics and/or about a particular language? A project such as the Greek and Latin Treebanks contain more than 1 million corrected syntactic annotations on individual words. How well can we detect patterns of difficulty among individual students and then use those patterns to personalize instruction?

Finally, we looked at the opportunities that corpora provide for students to make tangible contributions and then conduct their own research. Thus, students may begin by annotating data that either constitutes a stand-alone (but repurposable) corpus or augment existing annotation. Such annotations are, if well-executed, themselves tangible contributions to scholarship. Students can then conduct their own research based on these annotations and have an opportunity to conduct meaningful research and generate new knowledge that can be automatically linked to passages on which it sheds light.

Aside from the two formal workshops, the DFG/NEH Bilateral Project provided us with a framework to develop at least three major project outcomes: (1) an edited issue of the *Journal of Computing and Cultural Heritage*; (2) the NEH Working with Text in a Digital Age Institute for Advanced Technology in the Digital Humanities; (3) the Humboldt Professorship in Digital Humanities and associated Open Philology Project.

## Project Outcomes

### 1) April 2012 issue of the *Journal of Computing and Cultural Heritage*

The Tufts Workshop resulted in an edited issue of the ACM *Journal of Computing and Cultural Heritage* (JOCCH: http://jocch.acm.org/): Volume 5 Issue 1, April 2012 - http://dl.acm.org/citation.cfm?id=2160165&picked=prox

Gregory Crane and Anke Lüdeling, Introduction to the special issue on corpus and computational linguistics, philology, and the linguistic heritage of humanity.[3]

**Abstract**: The articles 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

[3] http://dl.acm.org/citation.cfm?id=2160166&CFID=230786826&CFTOKEN=25766783

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.

David Bamman and David Smith, Extracting two thousand years of latin from a million book library.[4]

**Abstract**: With the rise of large open digitization projects such as the Internet Archive and Google Books, we are witnessing an explosive growth in the number of source texts becoming available to researchers in historical languages. The Internet Archive alone contains over 27,014 texts catalogued as Latin, including classical prose and poetry written under the Roman Empire, ecclesiastical treatises from the Middle Ages, and dissertations from 19th-century Germany written—in Latin—on the philosophy of Hegel. At one billion words, this collection eclipses the extant corpus of Classical Latin by several orders of magnitude. In addition, the much larger collection of books in English, German, French, and other languages already scanned contains unknown numbers of translations for many Latin books, or parts of books. The sheer scale of this collection offers a broad vista of new research questions, and we focus here on both the opportunities and challenges of computing over such a large space of heterogeneous texts. The works in this massive collection do not constitute a finely curated (or much less balanced) corpus of Latin; it is, instead, simply *all* the Latin that can be extracted, and in its reach of twenty-one centuries (from approximately 200 BCE to 1922 CE) arguably spans the greatest historical distance of any major textual collection today. While we might hope that the size and historical reach of this collection can eventually offer insight into grand questions such as the evolution of a language over both time and space, we must contend as well with the noise inherent in a corpus that has been assembled with minimal human intervention.

David Mimno, Computational historiography: Data mining in a century of classics journals.[5]

**Abstract**: More than a century of modern Classical scholarship has created a vast archive of journal publications that is now becoming available online. Most of this work currently receives little, if any, attention. The collection is too large to be read by any single person and mostly not of sufficient interest to warrant traditional close reading. This article presents computational methods for identifying patterns and testing hypotheses about Classics as a field. Such tools can help organize large collections, introduce younger scholars to the history of the field, and act as a “survey,” identifying anomalies that can be explored using more traditional methods.

Hagen Hirschmann, Anke Lüdeling, Amir Zeldes, Measuring and coding language change: An evolving study in a multilayer corpus architecture.[6]

[4] http://dl.acm.org/citation.cfm?id=2160167&CFID=230786826&CFTOKEN=25766783
[5] http://dl.acm.org/citation.cfm?id=2160168&CFID=230786826&CFTOKEN=25766783
[6] http://dl.acm.org/citation.cfm?id=2160169&CFID=230786826&CFTOKEN=25766783

**Abstract**: Our article explores the possibilities of using deeply annotated, incrementally evolving comparable corpora for the study of language change, in this case for different stages from Old High German to New High German. Using the example of the evolution of German past tenses, we show how a variety of categories ranging from low to high complexity interact with the choice between competing linguistic variants. To adequately explore the influence of these categories, we use a multilayer corpus architecture that develops together with our study. We show that a combination of quantitative and qualitative analyses can recognize relevant contextual factors, which feed into the addition of new annotation layers applying to the same data. By making our categorizations explicit as corpus annotations and our data available to other researchers, we promote an open, extensible, and transparent mode of research, where both raw data and the inferential process are exposed to other researchers.

## 2. Working with text in a digital age: http://sites.tufts.edu/digitalagetext/

Anke Lüdeling and Gregory Crane, co-PIs on this DFG-NEH Bilateral Workshop, used their experiences to develop "Working with Text in a Digital Age", a three-week NEH Institute for Advanced Technology in the Digital Humanities (http://www.neh.gov/odh/). Tufts University hosted this workshop on July 23-August 10, 2012.

This institute combined traditional topics such as TEI markup with training in methods from Information Retrieval, Visualization, and Corpus and Computational Linguistics.
Co-directors were Monica Berti and Gregory Crane, Tufts University; Anke Lüdeling, Humboldt University. The institute also includes code with which to show participants how to develop a demo edition: https://github.com/TuftsUniversity/tei-digital-age.

## 3. Open Philology Project

In large measure because of the connections developed within this project, Gregory Crane was appointed an Alexander von Humboldt Professor of Digital Humanities at the University of Leipzig: http://www.humboldt-foundation.de/web/ahp-2013-en.html. This award brings with it 5,000,000 euros in support over five years. Professor Crane will use this to support the Open Philology Project, dedicated to developing new methods by which to analyze textual sources from the past.

The Humboldt Chair of Digital Humanities at the University of Leipzig sees in the rise of Digital Technologies an opportunity to re-assess and re-establish how the humanities can advance the understanding of the past and to support a dialogue among civilizations. Philology, which uses surviving linguistic sources to understand the past as deeply and broadly as possible, is central to these tasks, because languages, present and historical, are central to human culture. To advance this larger effort, the Humboldt Chair focuses upon enabling Greco-Roman culture to realize the fullest possible role in intellectual life. Greco-Roman culture is particularly significant because it contributed to both Europe and the Islamic world and the study of Greco-Roman culture and its influence thus entails Classical Arabic as well as Ancient Greek and Latin. The Humboldt Chair inaugurates an **Open Philology Project** with three complementary efforts that produce open philological data, educate a wide audience about historical languages, and integrate open philological data from many sources: the **Open Greek and Latin Project** organizes content (including translations into Classical Arabic and modern languages); **the Historical Language e-Learning Project** explores ways to support learning across barriers of language and culture as well as space and time; **the Scaife Digital Library** focuses on integrating cultural heritage sources available under open licenses. For further information: http://sites.tufts.edu/perseusupdates/2013/04/04/the-open-philology-project-and-humboldt-chair-of-digital-humanities-at-leipzig/

## Lessons Learned

The format of this project — two workshops — was fairly standard. Most of what we would put in a white paper are in the project outcomes (especially the JOCCH papers).

Innumerable issues emerge when organizing joint, international conferences and we are constantly reminded about the need to plan early and plan for the unexpected. The most important lesson for American researchers to learn is that in Germany the Humanities, Linguistics, and Computer Science are all simply aspects of *Wissenschaft*. There is no separate NSF and NEH — the DFG supports the Humanities, Social Sciences and Natural Sciences alike.

The general nature of Wissenschaft and the all-encompassing mission of the DFG provides an immense opportunity for US humanists collaborating with Germany. Where the NSF cannot regularly support Humanities research and US Computer Scientists have difficulty funding Digital Humanities research, this barrier does not exist in Germany. German Computer Scientists can collaborate with Classicists or Historians as easily as they can with Physicists or Chemists. Organizations such as the Humboldt Foundation pride themselves on serving all fields and on rewarding excellence, with no quotas for any discipline or field.

Digital Humanists in the US who wish to collaborate with Computer Scientists should look for German partners. The DFG/NEH Bilateral Program thus provides a unique opportunity for American Digital Humanists.