SBE2020: Analyzing human systems across time, space, language, and culture[1]

Alison Babeu, David Bamman, Lisa Cerrato, Gregory Crane, Rashmi Singhal
Perseus Project, Tufts University

Abstract: Due to the rise of very large, heterogeneous collections, increasingly sophisticated multilingual services, and expanding high performance computing infrastructure, we are now in a position to begin studying 4000 years of linguistic data from around the world, tracing change within languages, the interaction of languages, the evolution and circulation of ideas, and the patterns of human society. Language has been an impenetrable barrier – we can reach any point on the globe in a matter of hours but the amount of time required to master a new language remains unchanged. We can, however, now begin to work with far more languages than we could ever study, much less master. We are now in a position to pursue broader questions and to pursue these with greater rigor than would have been possible in print. A great deal of work remains to be done, however, for very large collections are not scientific corpora and need extensive processing, and many written sources do not yet lend themselves to optical character recognition. Simply scaling up existing systems to analyze millions of books poses software engineering challenges. Perhaps most important of all, we need to train a new generation of researchers who can bridge the intellectual gaps between the relevant computational methods and new research for social, behavioral and economic sciences.

1. Fundamental Question: Understanding 4,000 years of the human record

We need to understand the deep history of cultural systems if we are to understand their dynamics in the present. To do this we need to be able to work with the full linguistic record of humanity – 4,000 years of data from around the world representing thousands of linguistic systems. Few researchers are able to study the circulation and evolution of ideas from Greek to Arabic and then back to Europe via Latin translation – the barriers of language have been too challenging and the underlying sources are too scattered. But the Greek-Arabic-Latin pathway is only one element of the far more complex network of interacting languages, cultures, and ideas that produced the complex human systems in which we now live.

1) Problems of scale: even if we restrict ourselves to a single, technically tractable language such as Latin, we already have an open source collection of 1.7 billion words in digital form – two orders of magnitude larger than the 10 million word corpus of Latin preserved through c. 500 CE. The Munich-based Thesaurus

Linguae Latinae has been developing a new dictionary for the 10 million word classical corpus since 1894 and has published volumes through the letter O. With its staff of 20, the estimated completion time is c. 2050 – methods that clearly do not scale to the actual corpus of Latin preserved. Latin is only one of many historical languages for which data survives. The corpora of Sanskrit, Classical Chinese and Classical Arabic offer similar challenges as digital corpora increase in size. How do we visualize changes within individual languages? How do we visualize linguistic developments within a single language or the transformations of one language into another (e.g., Latin evolving into Romance languages)?

2) Problems of analogue sources: Historical sources are preserved in a wide range of written systems, on media such as stone, clay, metals of all kind, plaster, bone, papyrus, parchment, paper and marked both as inks and as engravings. Even supposedly 2d sources of linguistic data, such as printed books, often require 3d analysis because the source books cannot be opened flat, while other textual sources are written on curved surfaces or the writing itself is a 3d entity inscribed or pressed onto the written surface. Even hand-written sources in English on effectively flat surfaces in perfectly distinct ink poses serious challenges to current technologies. How do we visualize these varying sources consistently such that we foster, rather than impede, broader circulation of linguistic data?

3) Problems of historical change: Historical linguistics is one of the oldest fields within the academy. Latin evolves into languages such as Italian, French, Spanish, Portuguese, and Romanian. Classical Arabic shifts into the dialects that vary widely across thousands of miles. Access to growing corpora and increasingly sophisticated automated methods of analysis and visualization open up fundamentally new research pathways.

4) Problems of heterogeneity: In the approximately 4000 years of preserved linguistic data, Europe, Asia and Africa represent a single, extended network. We need to be able to trace two complementary classes of problem. On the one hand, languages evolve and interact: Latin spreads across the Mediterranean, Old French and Arabic pour into English and Persian, while Persian vocabulary spreads west into Egypt, north into Turkey, and south into the Indian subcontinent. At the same time, concepts lead across linguistic and cultural boundaries. These include not only topics from religion, politics, literature, and art but also scientific categories, data, and methods. How can we foster the interoperability needed to trace these two classes of phenomenon from Beijing to Birmingham?

5) The material record: Linguistic sources are intimately linked to the worlds of which they are a product, while the material record recovered provides data that not only augments but often conflicts with what we learn when linguistic sources survive. Researchers need scalable methods by which to integrate the increasingly large and heterogeneous datasets about the linguistic and material records of humanity.

2. Fundamental Science, Capacity, Infrastructure

Fundamental science

In this area at this point, we do not need fundamental science as much as we need fundamental sciences.

The questions that we pose here do not, at least yet, lend themselves to any one intellectual framework. Human systems are fundamentally interdisciplinary and the questions that surround them draw upon, but can only artificially be contained within, traditional disciplinary structures. When we work with four thousand years of linguistic data we involve – or should involve – every aspect of the social, behavioral and economic sciences. And we need to draw as well upon earth and biological sciences and methods from applied mathematics.

The challenge that we face is to develop new intellectual communities, which in turn create new intellectual structures from the resources already available. Logistical challenges have artificially increased the gap between the linguistic and material records – it has taken a great deal of legwork to assemble, much less visualize the results from scattered archaeological sites, while researchers have only been able to work directly with the linguistic data that they could personally analyze in the languages that they had studied. These barriers have, in turn, constrained the degree to which researchers in social, behavioral, and economic sciences could work with the human record.

Capacity

The development of appropriate capacity is challenging. Linguistic resources have evolved to meet the needs of print research: they reflect the loose structures of paper publication and are designed for human researchers with extensive expertise, much of it implicit and ill defined. We need to train a new generation of researchers who can both contribute to our understanding of human systems over time and also develop the infrastructure to support researchers from many disciplines.

On the one hand, many of the most talented students of historical languages and sources have traditionally focused their attentions on developing re-purposeable data and basic infrastructure. Editors organized textual data for use by a wider audience – a Greek edition of Euclid may have assumed a knowledge of Greek and of Mathematics but it also shielded the reader from the need to reconstruct the tangled manuscript sources for the *Elements*. Print dictionaries organized lexical information to expand intellectual access to particular languages, corpora and even authors. Commentaries summarized research results and questions about particular authors.

At the same time, these infrastructural research projects have fallen out of fashion in recent years – in effect, the research communities working with historical sources decided, in effect, that the infrastructure was good enough and focused more of its energy

on interpretative scholarship (largely designed to advance tenure, promotion and prestige within self-contained disciplinary networks). PhD programs in the humanities are not training the researchers that we need to support research that draws upon thousands of years of data and hundreds of languages.

We need to develop researchers in areas that include the following:

1) Corpus linguistics: Some languages have been the object of study for thousands of years but our linguistic resources are often prescriptive (i.e., grammars that portray an ideal of Ciceronian Latin) and, where they do describe, resort to generalizations (e.g., "common in Tragedy", "rare in good Latin prose"). Many of our linguistic sources contain normalized versions, excluding variations that cast light either upon the original version (in cases where there was a single original version) or upon the development of the language. We need a new generation of philologists who integrate the methods of corpus linguistics, developing systematic linguistic annotation as the foundation for their conclusions and creating machine actionable corpora that can support a wide range of research projects.

2) Computational linguistics: We have far more data than we can explore, much less analyze, by manual means. We need to work with billions, if not trillions, of words. We need to exploit automatic methods that are scalable and provide results of readily measured accuracy. These methods must address multiple languages – the vernacular languages of Europe co-evolved in conjunction with each other and with the changing forms of Latin that dominated education and formal publication for two thousand years.

3) Information retrieval: We need to visualize ideas as they flow within and across languages in time and space, much like currents of air or water. We need to be able to identify different networks as they emerge over time, with some ideas moving slowly from the Pacific to the Atlantic, others swirling in much tighter networks around the Mediterranean or the Indian subcontinent. We need multilingual information retrieval services that can detect ideas across hundreds of languages and then provide the analytical services so that researchers can probe into the primary sources themselves.

4) Cross-cultural expertise: We need to foster new intellectual communities and to create new configurations of disciplines but we also need to create intellectual communities that span stubborn language and cultural barriers. The University of Cairo, for example, supports a vigorous community of researchers with expertise in Greek and Latin who have the unique ability to analyze the circulation of scientific ideas as they moved from Greek into Arabic from 800 to 1000 CE and then from Arabic into Latin c. 1200. Researchers at the Universidad Nacional de Córdoba (Argentina), Universidad Pontificia Bolivariana in Medellín (Colombia), and elsewhere likewise have a unique perspective on the Colonial Latin that developed in the New World under the influence of Jesuit, Dominican and

Franciscan missionaries, whose profound influence on shaping the culture of an entire continent can still be felt today. We need to develop a generation of researchers who bridge such gaps, able to connect research in Arabic and English, English and Chinese, and other combinations.

Research that addresses these questions will also require sustained and significant *infrastructure* development:

1) An extensible network of Open Content linguistic corpora: These should provide increasingly systematic coverage of human languages and linguistic genres and must be in formats that are increasingly interoperable. Coverage should include particularly significant languages of well-defined communities (e.g., Hieroglyphic Egyptian and Hebrew) and the great *linguae francae* that speakers of many languages have shared and within which ideas have circulated and evolved for long periods of time (e.g., Sumerian, Akkadian, Classical Chinese, Sanskrit, Classical Greek, Latin, Classical Arabic, Persian).

   We can already see pragmatic, layered collections emerging that include some linguistic sources as 2d or 3d datasets (e.g., manuscripts and inscriptions), others as text automatically generated from images by Optical Character Recognition services, some as carefully curated collections with manually reviewed XML markup, and still others with labor-intensive forms of linguistic annotation (e.g., syntactic databases such as Treebanks).

2) A wide range of linguistic services: Such services either already exist or could be built, and those services already available need to be expanded and improved. The greatest challenge is to create services that can scale to very large collections and that are smart enough to grow more effective by interacting with a wide range of user communities.

Additional Reading

Abney, S. and S. Bird (2010, July). The human language project: Building a universal corpus of the world's languages. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, pp. 88-97. Association for Computational Linguistics. http://www.aclweb.org/anthology-new/P/P10/P10-1010.pdf

Gregory Crane, Alison Babeu, David Bamman, Lisa Cerrato, Rashmi Singhal. Tools for Thinking: ePhilology and Cyberinfrastructure. In Working Together of Apart: Promoting the Next Generation of Digital Scholarship: Report of a Workshop Cosponsored by the Council on Library and Information Resources and The National Endowment for the Humanities , , Washington, D. C. United States : Co-Sponsored by: Council on Library and Information Resources National Endowment for the Humanities, 2009-03 http://www.clir.org/activities/digitalscholar2/crane11_11.pdf

Lüdeling, A. and A. Zeldes (2007). Three views on corpora: Corpus linguistics,literary computing, and computational linguistics. Jahrbuch für Computerphilologie 9, 149-178. http://computerphilologie.tu-darmstadt.de/jg07/luedzeldes.html