Digital Linguistics on the Web: Some Personal Reflections

Brian D. Joseph
The Ohio State University
Schloss Dagstuhl Seminar on Computational Humanities
July 2014
The internet/web runs on ....

- advertising??

  (NB: could Google or Facebook exist without advertising?)

- commercial activity??

  (cf. Amazon.com, ebay.com, ...)

or ...

• pornography??

... or ...


... LANGUAGE

• even if a picture is worth a thousand words, and even if images vastly outnumber text-“chunks” (even with 18,000,000,000 words in all Wikipedias alone), and even if we do not count advertising pages and porno sites with their preponderance of images, could we have an image-only world-wide web?

• Language seems indispensible for the internet/web to work and to run efficiently (at least for search-engine-optimization and for accessibility issues)
• language??, or rather, languages??

NB: • lots of languages represented on the web besides English and other “big” languages; note there are c. 300 different language editions for Wikipedia

• Internet % of English was c. 80% a few years ago but is declining (now c. 55%), even as number of English users increases
LINGUISTICS = SCIENTIFIC/OBJECTIVE study of language as a STRUCTURED COMMUNICATION SYSTEM, as a TOOL FOR SOCIAL INTERACTION, as an EXPRESSION OF HUMAN IDENTITY, as a CHANGEABLE YET STABLE OBJECT OF HUMAN CONSTRUCTION, and as a CONVEYOR OF HUMAN CULTURE
What *could* web mean for language, for language users, and for linguists?

- a place for public and private use of a language (cf. Albert de Swaan and “language as a hypercollective public good”)

- a place for human expression through language

- a place for data to reside (language as a museum object, language as a monument, language as living heritage)

- a place where data is created (cf. Christiane’s Seminar paper and “culturnomics”; studies of internet usage; etc.)
During the seminar so far, we have already seen lots of excellent examples of linguistics and language-related intersections of computational methods / digitalizations with humanistic questions, thanks to the many computational linguists here.
Some amazing language-crossed-with-digital-humanities sites that I have nothing to do with, except as a user:

- TITUS Project ([http://titus.fkidg1.uni-frankfurt.de/indexe.htm](http://titus.fkidg1.uni-frankfurt.de/indexe.htm); cf. Jost Gippert’s baby)

- LinguistList ([www.linguistlist.org](http://www.linguistlist.org))
Some web-based projects that I have been involved in that in some small way reflect the range of ways in which Digital Humanities is manifested in the field of Linguistics:

- U of Texas Linguistics Research Center, as co-author of Albanian on-line materials:

  [http://www.utexas.edu/cola/centers/lrc/eieol/albol-0-X.html](http://www.utexas.edu/cola/centers/lrc/eieol/albol-0-X.html)
• ELCat (Catalogue of Endangered Languages), produced by the University of Hawai’i at Mānoa and The Institute for Language Information and Technology (The LINGUIST List), with funding from NSF and supported by a team of global experts (BDJ as one of the “experts”), the knowledge basis for The Endangered Language Project (ELP, supported by Google):

ELP:  http://www.endangeredlanguages.com/

ELCat: http://linguistlist.org/projects/elcat.cfm
• Duke University Slavic and Eurasian Language Resource Center (SEELRC), as member of Advisory Board, instructor in Summer School, author in web-journal *Glossos*)

http://slaviccenters.duke.edu/
• OSU Linguistics: [http://linguistics.osu.edu/research](http://linguistics.osu.edu/research) (as colleague)

  --Buckeye Speech Corpus: [buckeyecorpus.osu.edu/](buckeyecorpus.osu.edu/)

  --OhioSpeaks: [ling.osu.edu/ohiospeaks/index.html](ling.osu.edu/ohiospeaks/index.html)

  --BLN COSI Pod: [bln.osu.edu/LanguagePod.php](bln.osu.edu/LanguagePod.php)

  --Herodotos (cf. Brown & Joseph 2014ab)
Where/What else? (with me as spectator!):

• various documentation projects:

  --The Archive of the Indigenous Languages of Latin America (AILLA), at U of Texas

  [http://www.aiilla.utexas.org/site/welcome.html](http://www.aiilla.utexas.org/site/welcome.html)

  --Survey of California and Other Indian Languages and California Language Archive, at U Cal Berkeley

  [www.linguistics.berkeley.edu/Survey/](http://www.linguistics.berkeley.edu/Survey/)

  [http://cla.berkeley.edu/](http://cla.berkeley.edu/)
• linguist-run sites/projects that go beyond just language:

  • Bill Davies (U of Iowa) and Madurese traditional story-telling (A digital archive of Madurese folk tales and historical narratives) with videos:

    http://digital.lib.uiowa.edu/madurese/

  • Daryl Baldwin (Miami Univ., Ohio) and Myaamia Center (documenting Miami Indian life, culture, history, language):

    https://www.miami.muohio.edu/about-miami/diversity/miami-tribe.html
Blogs and fora that are linguistically well informed, thus unlike many language-related sites with, e.g. grammar rants):

• Language Log:
  
  http://languagelog.ldc.upenn.edu/nll/

• Neal Whitman (OSU Ph.D., 2002) and his blogs:
  
  http://literalminded.wordpress.com/

NB: the comment areas on these utilize computational means not available in, e.g., print media, and so are (sort of) digital linguistics
And, not-Linguistics on the web:


(with a silly statement about language change being so rapid now with social media and such – yet, how to measure rate of language change is a nontrivial question, not to be decided impressionistically based on the fates of a few lexical items)
And lots and lots and lots of other language-related objects on the web ... (think YouTube and sound files alone, for instance) ...