GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

CLT

Towards a knowledge-based culturomics

Lars Borin

Språkbanken • Swedish Language • University of Gothenburg

kickoff meeting • 31st Jan 2013

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

"culturomics"

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel,[1,2,3,4,5]*† Yuan Kui Shen,[2,6,7] Aviva Presser Aiden,[2,6,8] Adrian Veres,[2,6,9] Matthew K. Gray,[10] The Google Books Team,[10] Joseph P. Pickett,[11] Dale Hoiberg,[12] Dan Clancy,[10] Peter Norvig,[10] Jon Orwant,[10] Steven Pinker,[5] Martin A. Nowak,[1,13,14] Erez Lieberman Aiden[1,2,6,14,15,16,17]*†

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics,' focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

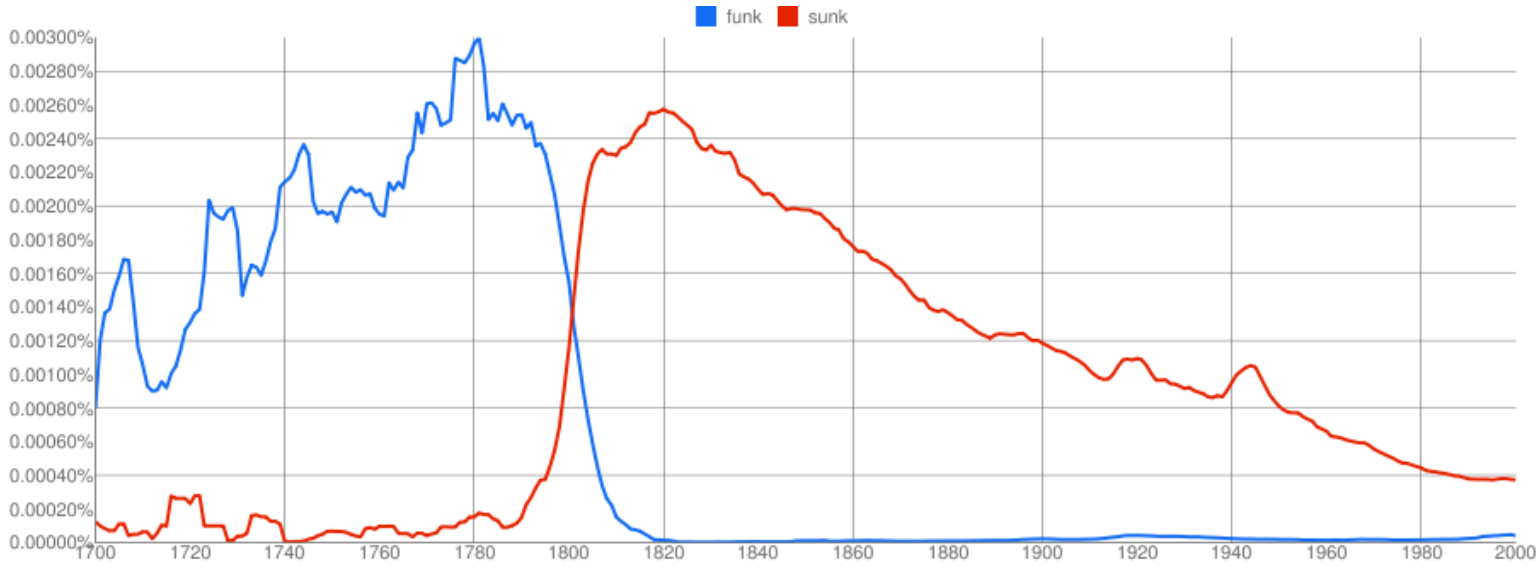
(*Science* 331, 176 (2011))

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

knowing your data, 1



("funk" vs. "sunk" in Google n-grams)

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

knowing your data, 2

Antal träffar: 781

Föregående 1 2 3 4 5 6 7 8 9 10 11 .. 31 32 Nästa Visa kontext

	DALPILEN	
i Björsberg kl. 4 em. och i gamla c.is,ionftrgkand dat	8ven	Skall jö.
	8ven	» l<t 0llw.
	8ven	»!l tolk fäljl!!
Söndagen » kollekt:	8ven	» ka foli.
	8ven	Johan LinditrSm eoh B««nhiid E isabat Bngitrflm, bida fr. Korania
att je,nle det trvelclrideton s,>rtsarande i lieflerin^ » ser,nen sördeliällog	8ven	» Ka l '» dket.
Kollekt till	8ven	» ka Bofiafdrsamlingen i Paris.
> ,,,,,	8ven	» K l,nr,°lKIl>la as <!!.
!>>ulites,	8ven	» l<.i, 1') sliir ne !> ^' rnn^li.r l) m«Il>^ »- Ol » s ,', m<' !I<' i '.
2 mecl unclerlagcl »	8ven	» Ka nrä 8amt upplvsancle text, ul^ilvet al ^?.
SMagm kolllkt:	8ven	' k »; kyrkans Diako nittyreleei lånefond.
	8ven	Johan Salén från Helsingborg, 64 år, 5 rn, 21 d.
	8ven	" Bo: i
	8ven	Holger, son till Edvard Be ' glund, Fortby.
Apri: S. rjiurid Gullan Viktoria, dotter till	8ven	Gerhard Soneaton, i kålf1, 8 mån
	8ven	Olof, gon till 8ven Olof Öländer i Envikabyn.
8ven Olof, gon till	8ven	Olof Öländer i Envikabyn.
navalet i Östra härad, Blekinge, hvilket i tisdags egde rum, återvaldes hemmansegaren	8ven	Aruoldssou.
) u » oeb mörk, oräinär oob	8ven	» K;
	8ven	kl ?8il2ll<ier,
Hs	8ven	» K » legeringen Kelu !lmäKtig » s Dtvansrings-^gent.
Skrftdderitillskäraien,	8ven	Johan Nilsson-Neijbert fr. Vagnmakaren 3, 69 år.

(instances of “8ven” i the oldest Digidaily material)

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

the hitch (Mark Liberman, LL 17 Dec, 2010)

There are also a large number of cases where you'd like to group word-strings into categories: dates, organizations, minerals, place names, novelists, etc., and then treat these categories (rather than words or word strings) as units of analysis. Again, there are well-known techniques for inducing such categories in text collections — but to use these techniques, you need to be able to have the text collection in hand so as to be able to run your algorithms over it.

Many — maybe most — questions about historical texts are like these last few examples: relatively easy to answer if you have a corpus in hand, and not addressed very well (if at all) by a collection of "culturomic trajectories", defined as the year-by-year time-functions of common word sequences. In particular, nearly all questions about the history of the English language fall beyond the grasp of time-functions of n-gram frequencies. This is not to deny the interest and value of such time functions. It's just that they're not nearly enough.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

project consortium • strengths

- Chalmers • machine learning, data mining
- Lund • semantic parsing and role labeling
- Gothenburg • language resources, knowledge-based LT

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

our proposal

- general:
  - work on Swedish (and smaller datasets)
  - apply deep linguistic processing to texts
  - combine knowledge-rich and statistical approaches

- specifics:
  - a Swedish Watson
  - discover and track semantic change over time

why we got funded

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Overall grade for the application 6

(7=Outstanding, 6=Excellent, 5=Very good to excellent, 4=Very good, 3=Good, 2=Weak, 1=Poor)

Motivation for the overall assessment of the application

The panel concluded that this was a highly technical project. The reputations and experience of the team seem matched to the challenge although even this could be a little difficult to judge without more detail as to what will actually be achieved – what success for the challenge is defined to be.

The project therefore would gain by specifying very clear goals or outcomes against which its worth can be considered and its achievements judged. The scientific reviewers agreed that this project falls short on the detail with which it specifies these aspirations. However, the panel were inclinded to be supportive to the project overall.

It was noted that in the recent past, Culturomics research was able to retroactively predict the 2011 Arab Spring and successfully estimate the final location of Osama Bin Laden to within 124 miles. This is very impressive, since social science is not ususlaly so precise. It was agreed that finding ourselves in a situation of text-information-overload, more and better applications of semantic processing are in great need. Focusing on the rich treasure of Swedish texts seems adequate and highly relevant for this grant. The researchers also draw from long-standing concepts of linguistics and language technology, which gives it a solid disciplinary grounding. If the researchers achieve to create concrete and user-friendly tool that can be applied by others, the outcomes of this project could have a very high and real impact.

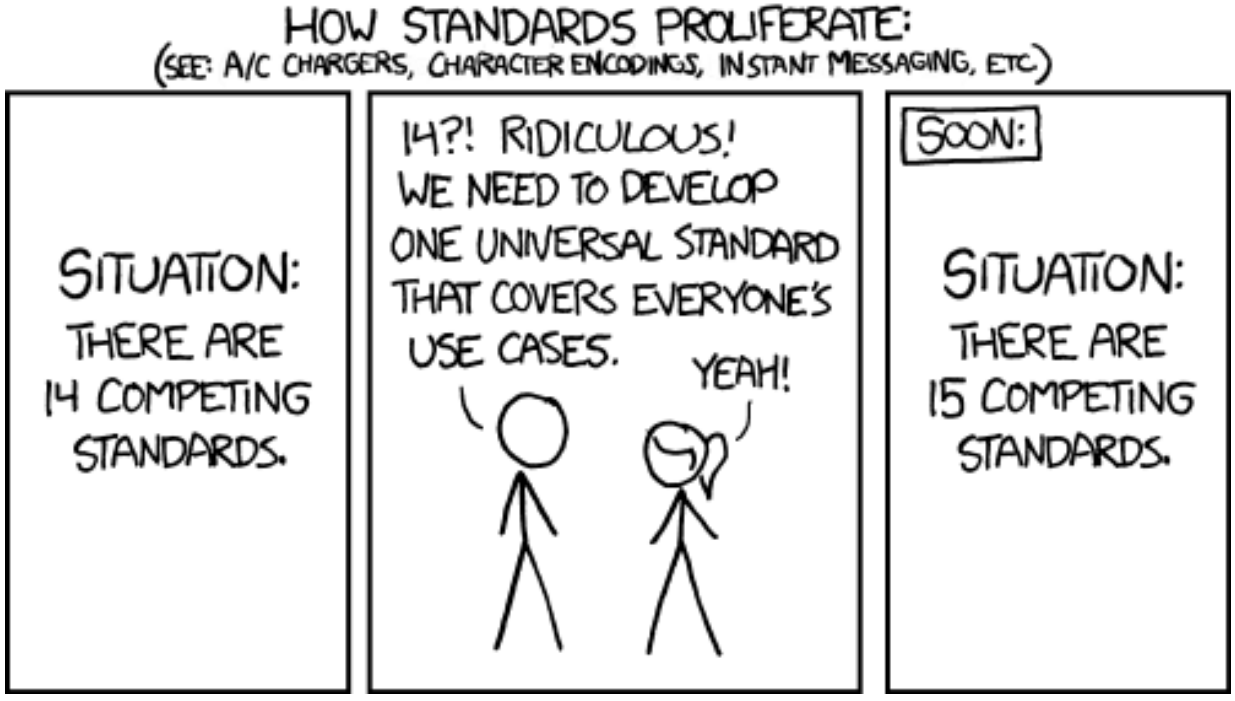
GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

challenges

- operationalizing and concretizing our goals
- coordinating our work:
  - in time
  - in objectives
  - in teroperability:



(<http://http://xkcd.com/927/>)