Using Language in Behavioural Science

Daniel Sgroi
University of Warwick, CAGE & IZA
daniel.sgroi@warwick.ac.uk

25 May 2020

Using Text in Behavioural Science

- Why use text/language data?
- My focus today will be on two good reasons:
  - Answer the unanswerable ("Historical Analysis of National Subjective Wellbeing using Millions of Digitized Books").
  - Find the *right* answer ("Cultural Identity and Social Capital in Italy").
- But in practice there are many more good reasons.

Historical Analysis of National Subjective Wellbeing using Millions of Digitized Books

Using language to find an answer...

The Paper

- "Historical Analysis of National Subjective Wellbeing using Millions of Digitized Books", Nature: Human Behaviour 15 October 2019, joint with Thomas Hills, Eugenio Proto and Chanuki Seresinhe.
- The focus is on developing a measure for national happiness thatwithout language data would be close to impossible.
- Our primary objective was to produce a workable proxy for subjective wellbeing going back to 1800, which would enable direct comparisons with GDP over that period.
- But survey data simply isn't available and lab data would inappropriate so we need to be innovative...
- The approach was to infer mood from text. Our methods rely on the digitization of books and newspapers, available in numerous corpora, such as the Google Books corpus, the British Newspaper project and the COHA corpora.

Language Corpora

- To make progress we needed both a corpus of language (a source of text data) and a set of word norms (what individual words tell us about mood).
  - *Google Ngrams* (`https://books.google.com/ngrams`) based on a digitized database of several million published books. We focus on data for 4 languages, English (British), English (American), German and Italian.
  - "Find My Past" data from the British Library's "British Newspaper Project" which covers 65 million newspaper and periodical articles from the UK across 200 periodicals going back to 1710.
  - US English COHA Corpora which includes 400 million words from 1810-2000.
  - 2 sentiment indices: a "National Pleasantness Index" and "National Polarity Index" derived from SenticNet data.
- Our results were robust to the choice of corpora in most of what follows we focus on the Google corpus.

Word Norms

- Word valence rating norms ask participants to rate each word from a list on how positive or negative they perceive a word to be.
- To allow for comparison across languages, all of our valence norms use a subset of words. There is a list of a thousand words that served as the basis for developing valence ratings for multiple languages through several independent studies.
- For English, we use ANEW which contains about 10,000 words rated on a 1 to 9 valence scale by a group of subjects.
- For German, we used the Affective norms for German sentiment terms. This is a list of 1003 words, a German translations of the ANEW list. The valence ratings were collected on a -3 to +3 scale. The mean values were adjusted to reflect a 1 to 9 scale. For Italian, we used an adaptation of the ANEW norms containing 1121 Italian words, based on the ANEW material on a 1 to 9 scale.

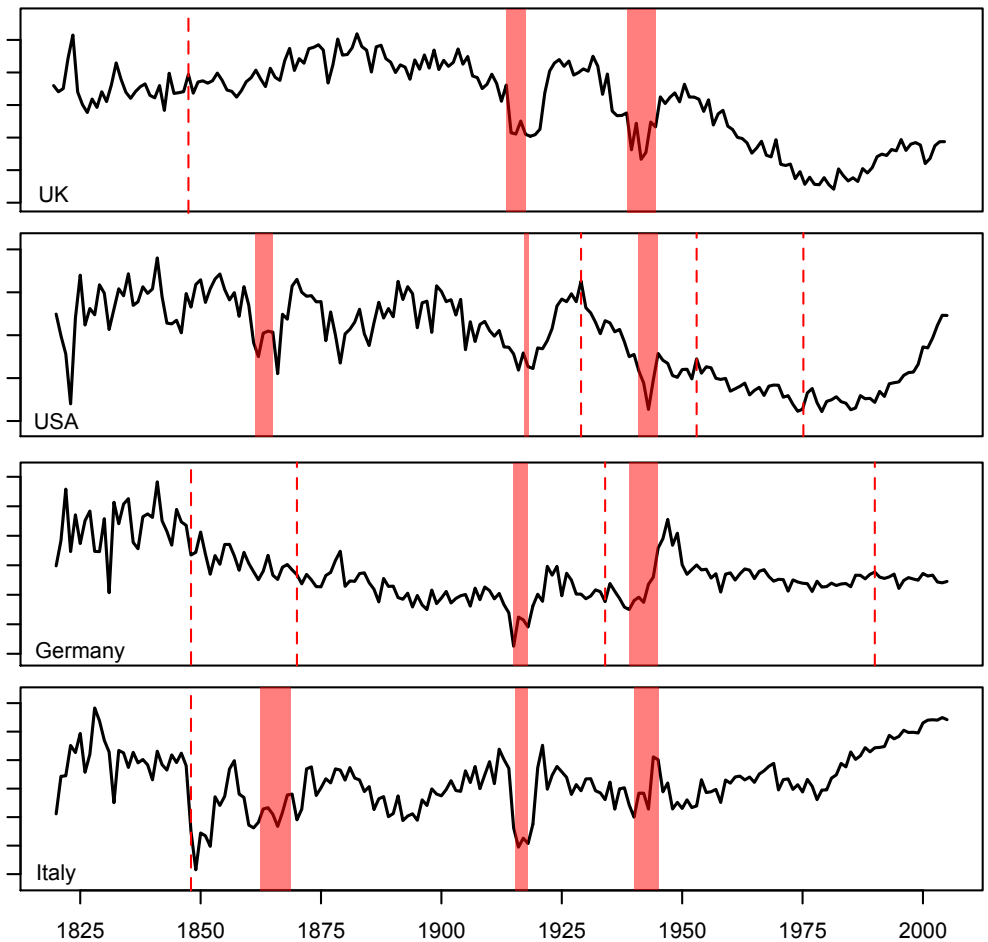
Valence and Words in Different Languages

- Some example valence ratings from ANEW:
  - High end: Happiness 8.53, Enjoyment 8.37, Vacation 8.53, Joy 8.21, Relaxing 8.19, Peaceful 8, Lovemaking 7.95, Celebrate 7.84.
  - Low end: Murder 1.48, Abuse 1.53, Die 1.67, Disease 1.68, Starvation 1.72, Stress 1.79, Unhappy 1.84, Hateful 1.9.
- For each language we compute the weighted valence score, $Valence_t$, for each year, t, using the valence, v for each word, j, as follows,

$$Val_{i,t} = \sum_{j=1}^{n} v_{j,i} p_{j,i,t};$$

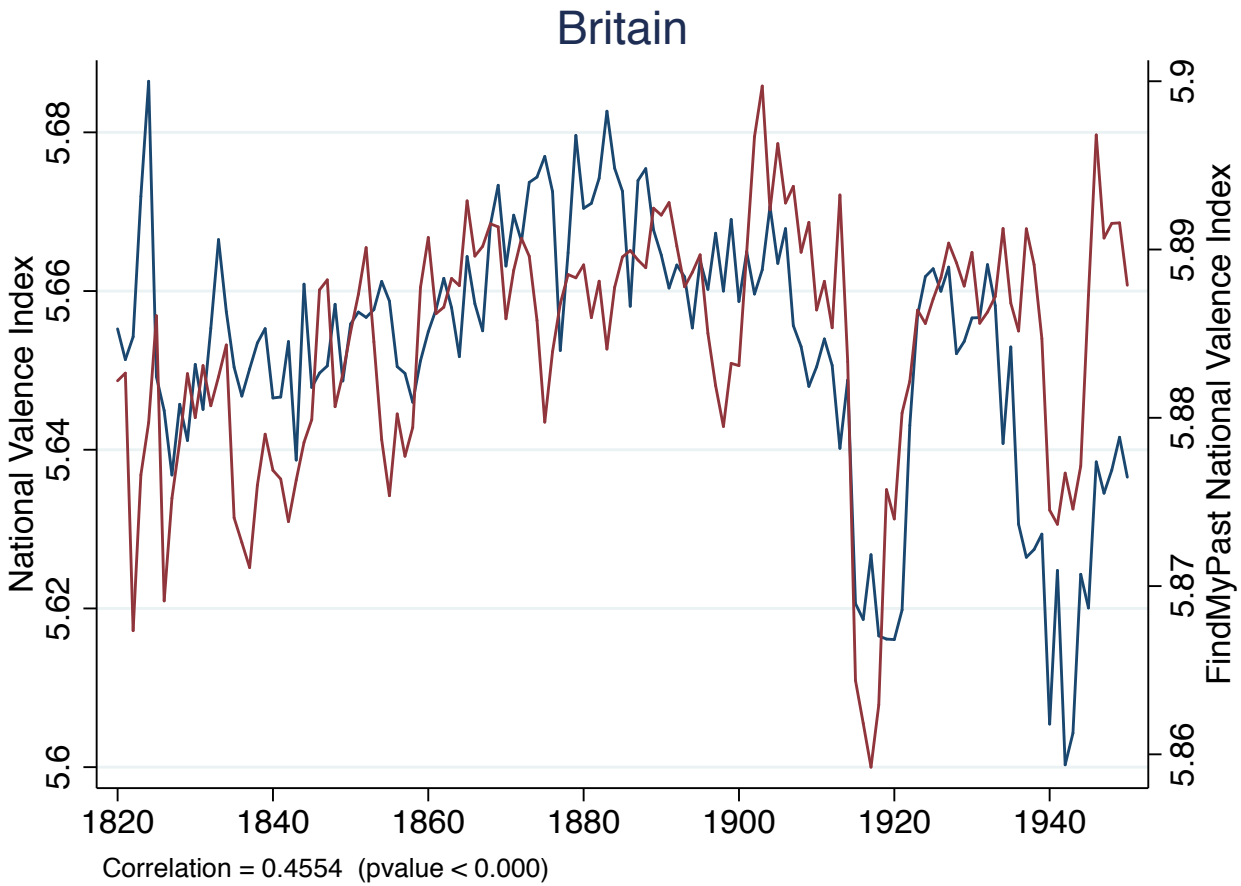
- Note that $v_{j,i}$ is the valence for word j as found in the appropriate valence norms for language i, and $p_{j,i,t}$ is the proportion of word j in year t for the language i.

A Time-Series Plot of the NVI, 1820-2009



Comparing Newspapers and Books for the UK, 1820-1950

Note: Blue is the book-based NVI, red is based on newspapers.



Summary of the Main Findings

- Our index based on average word valence of a language predicts country aggregate subjective well-being for several countries and correlates well with survey-based measures when they exist.
- But more than that it can go back much further than existing measures that use non-text data.
- Our index correlates positively with life expectancy, GDP (mildly) and negatively with conflict.
- Our findings are robust to different corpora (books, newspapers) and word norms and to the stability of word meanings over time by looking at word neighbourhoods.
- Without language data we would still be without any long-run time-series measure of national happiness.

Cultural Identity and Social Capital in Italy

Using language to find the *right* answer...

The Paper

- "Cultural Identity and Social Capital in Italy", WIP with Emanuele Bracco, Federica Liberini, Ben Lockwood, Francesco Porcelli and Michela Redoano.
- The aim is to try to understand the relationship between identity and social capital (trust in others and institutions).
- Special focus on Italy: one nation only relatively recently, significant regional variation in social capital (OpenCivitas), and in language.
- Normal survey measures that link identity to social capital are subjective and possibly endogenous, so we use language as a more objective measure of identity.

Social Capital

- Past research has pointed to the importance of cultural identity in establishing beliefs and customs.
- Much of this literature is based on studies of migration across countries, especially the "melting pot" of cultural mixing that took place during the multiple waves of immigration into the USA and share a common theme: that individuals from different cultural background can and do make different choices even when they share the same current environment.
- In this paper we argue that certain key aspects of cultural identity may take many generations to become established.
- In particular we study the powerful influence that family background plays in establishing patterns of cooperation, reciprocity and trust between citizens and government or society.
- The latter concept is normally referred to as social capital.

Italy

- The key method in our analysis is a combined survey and experiment administered by Qualtrics through their online panel in April 2019.
- In total 1500 subjects took part, 500 from each of three major Italian cities: Milan, Turin and Rome.
- The participant population was pre-screened to admit only subjects with both sets of grandparents of Italian origin, but otherwise was pre-selected by Qualtrics to ensure a demographic spread that resembles the wider Italian population.
- The study which took approximately 20 minutes, and since the study included (incentivized) experimental elements it was registered in advance in the AEA RCT Registry.
- The base earnings for each subject was 3 euros, with bonus payments of up to 50 euros for each of the games.

Design

- First, participants were asked a set of questions about the their origins and their family's origins.
- Second, they were asked a series of questions designed to generate a language (and diet)-based measure of underlying cultural identity.
- Third, they played two incentivized games: a public goods game and a simple test of honesty (self-declared number of heads from 10 coin flips)
- Finally, they were asked a series of questions designed to measure their level of social capital.

Focus on Language

- Words in Italian dialect can be very different.
- For instance the word for towel (*asciugamano*) which in dialect could be *tuvagghia*, *sciugaman* or *macrame* depending upon the region of Italy.
- Each participant was asked to listen to four recorded sentences in the regional dialect of his or her grandparents.
- They were asked to translate the sentence and comment on their understanding.
- They were then asked to decide which of a series of words in dialect were most used by their own family.
- Next they were asked to
  - translate various regional sayings (assigned based on their parents' place of birth)
  - select which word they would use to describe their loved ones
  - state which word they might use to describe a melon (having seen a photo), a fruit which has many different names in different dialects.

PCA

- First we removed "neutral" answers which are not regional, then we awarded "1" for regional answers from the North, "2" for answers from the Centre and "3" for answers from the "South".
- This allowed us to say that the higher is the score for each individual, the more southern is their linguistic preference or understanding.
- We then carried out a simple PCA allowing us ot focus on the core questions that mattered and apply appropriate loadings.
- We then renormalized to a score of between 0 and 1.

Ranking Family Members

Outcomes on Different Measures of Identity

Dep. Var.:	Public Good (own contributions)		Public Good (partner contributions)	
	(1)	(2)	(3)	(4)
Geographical Identity of Participant:				
Objective Index	-1.719**		-1.408*	
	(0.0325)		(0.0971)	
Subjective Index		0.174		0.257
		(0.760)		(0.572)
People not helpful Dummy	-0.327	-0.554	-0.418	-0.825
	(0.526)	(0.328)	(0.379)	(0.270)
People not honest Dummy	-0.533	-0.408	-0.256	-0.213
	(0.212)	(0.424)	(0.579)	(0.676)
People not trustful Dummy	0.772	1.307*	0.629	1.199**
	(0.121)	(0.0985)	(0.251)	(0.0203)
Town of Residence				
Rome	-0.443**	-0.709**	-0.220	-0.503*
	(0.0248)	(0.0378)	(0.212)	(0.0584)
Turin	-0.531**	-0.866***	-0.238**	-0.470***
	(0.0206)	(0.00829)	(0.0346)	(0.00235)
Observations	1,497	1,192	1,497	1,192
R-squared	0.065	0.066	0.041	0.046

The Importance of Language

- Our bottom line results show that attitudes towards social capital do indeed relate to underlying identity and can take at least 2 generations to change.
- This only becomes apparent if you use language to measure identity: without this we would not see the link.
- Language is therefore essential in making the link between behaviour and identity: but also seems to be a central part of who we are.
- We also validate the measure through dietary choices.

Conclusions

- Much of my recent work involves language and text analysis including more conventional lab experiments.
- For example:
  - How to convince people to behave in a more environmentally friendly fashion that uses text analysis to check which nudges work best and why some don't.
  - Why people behave the way they do while playing games in the lab: using text analysis to allow us to select among competing theories.
  - Exploring the interaction between language and personality.
- Bottom line: there are a huge variety of ways to use text analysis: answering questions that could not otherwise be answered, validating or supporting traditional methods, aiding understanding, and so much more...
- A very promising area for future work in behavioural science.