**ONLINE APPENDIX BJPS POVERTY PAPER**

**DICTIONARY WORD LISTS**

We use online dictionaries to identify the words in our major theoretical categories and try to include all words unless there are ambiguous meanings. (For example, the English word “society” almost always refers to high society in the eighteenth and early nineteenth centuries; only later does it take on the social connotations that it has today.) We then translate the words and try to use words that were commonly used in centuries past. We are native speakers in English and French and one of us has very high fluency in Danish. We have extensively read fiction from the eighteenth and nineteenth centuries in our languages. Finally, we look at how individual words perform in the snippets and choose words that have higher frequencies.

\* Poverty words around which snippets are built include:

poverty, pauper, destitute, indigent, needy, beggar, mendicant
(poverti/pauper/destitut/indig/needi/beggar/mendic)

fattigdom fattiglem trængende armod nødlidende tigger betler

fattigdom/fattiglem/træng/armod/nødlid/tig/betl

Pauvreté appauvri dépourvu indigent nécessiteux mendicant gueux

pauvret/appauvr/dépourvu/indigent/nécessit/mendi/gueux

\* Skills words include:

skill, ability, capacities, competence, qualification
skill/abil/capac/compet/qualif

Færdigheder evne kapaciteter dygtihed kvalifikation
Færd/evn/kapacit/dygt/kvalifikation

habileté aptitude capacités compétence qualification
habilet/aptitud/capac/compétent/qualif

\* Charity words include:

charity, benevolence, philanthropic, and beneficence
chariti/benevol/philanthrop/benefici

velgøren godgørenhed filantropiske næstekærlighed
velgør/godgør/filantropisk/næstekær

Charité bienveillance philanthropique bienfaisance
charit/bienveil/philanthrop/bienfais

\* Society words include:

England, English, Britain, country, folk, people, collective, communal, mutual, custom, social
england/english/britain/countri/folk/peopl/collectiv/communal/mutual/custom/social

Danmark Dannemark Dansk land folk mennesker kollektive fælles gensidig skik social
danmark/dannemark/dansk/land/folk/mennesk/kollektiv/fæl/gensid/skik/social

France (français) gen peuple collective commune mutuel coutume social
franc/pay/gen/peupl/collect/commun/mutuel/coutum/social

\* Individualism words include:

individual, independent, person, character, liberal, self
individu/independ/person/charact/liberal/self

individual uafhængig person karakter liberal selv
individuel/uafhæng/person/karak/liberal/selv

individuel indépendant person caractère liberal soi
individu/indépend/person/caractère/libéral/soi

\* Family words include:

Family marriage children parent mother father
famili/marriag/children/parent/mother/father

familie ægteskab børn forældre mor far
famili/ægteskab/børn/forældr/mor/far

famille mariage enfants parents mère père
famill/mariag/enfant/parent/mer/per

\* Government words include:

nation, government, law, council, commission, power, public, legal, committee, king, kingdom, emperor, crown, throne, municipal, parish, illegal, court
nation/govern/law/council/commiss/power/public/legal/committe/king/kingdom/emperor/crown/thron/municip/parish/illeg/court

stat, regering, lov, råd, commission, magt, offentlig, juridske, udvalg, kong, rige, kejser, krone, trone, kommune, sogn, ulovlig, ret
stat/regering/lov/råd/kommission/magt/offent/juridisk/udvalg/kong/rig/kejs/kron/tron/kommun/sogn/ulov/ret

etat, gouvernement, loi, conseil, commission, pouvoir, public, juridique, comité, roi, royaume, empereur, couronne, trône, municipal, paroisse, illegal, tribunal etat/gouvern/loi/conseil/commiss/pouvoir/public/jurid/comit/roi/royaum/empereur/couron/trôn/municipal/paroiss/illégal/tribunal

\* Church words include:

religious, rectory, church, cathedral, dissenter, Anglican, priest, God, spirit, bishop
religi/rectori/church/cathedral/dissenter/anglican/priest/God/spirit/bishop

religiøs, præstegård, kirke, domkirke, pietistiske, lutherske, præst, Gud, aand, biskop
religiøs/præstegård/kirk/domkirk/pietistisk/luthersk/præst/Gud/aand/biskop

religieux, presbytère, église, cathedral, dissident, catholique, prêtre, Dieu, esprit, évêque
religion/presbyter/églis/cathedral/dissident/cathol/prêtr/Dieu/esprit/évêqu

**OUR CHOICE IN COMPUTATIONAL TEXT ANALYSIS METHODS**

We calculate frequencies of poverty words appearing in the entire corpus for each country and the frequencies of words associated with our theoretical concepts within snippets of text surrounding our labor words. To calculate the word frequencies, we stem the words in the corpora and take out the stop words. We construct snippets of 50 words around all of the poverty words. The snippets are stemmed and filtered (i.e. stop words are taken out). The findings are not affected when the snippets include stop words, nor are the findings significantly affected if we choose snippets of 200 words instead of 50 words. We use difference of proportion tests to evaluate statistical significance.

Scholars using machine learning or natural language processing analyses in political science typically either use supervised learning techniques or unsupervised learning techniques. Supervised learning techniques may be used to classify works by organizing them into categories or to scale works by determining the position of a document on a linear scale, i.e. from left to right. For example, Hopkins and King (2010) develop a “bag of words” technique with which they hand-code texts into set of exhaustive categories and use the output to estimate the proportion of documents into each category. Their goal is to get an accurate estimate of document categories and to represent unstructured text as structured variables that can be analyzed statistically.

There are some methodological problems with supervised learning techniques. If the labeled set is not a random sample from the population, then it may not reflect the larger population. Moreover, it is very hard to find the true model (Hopkins and King 2010). Thus scholars also use unsupervised-learning techniques such as probabilistic topic models. Rather than beginning with keywords, one may ask the computer to identify the main themes or topics in a collection of documents. A topic is a “distribution over a fixed vocabulary,” and documents have multiple topics, each with their own particular distribution of topics (Blei). There are probabilities of topics appearing in each document and of words appearing in each topic. With processes such as the latent Dirichlet allocation if one simply specifies the number of topics in advance, one can use a “topic model algorithm” to infer the hidden topic structure, and compute the distribution of topics that best captures the collection of words in a document. Machine learning defines a joint probability distribution over observed and hidden random variables. So one computes the conditional distribution of the hidden variables, given the distribution of the observed variables. This conditional distribution is the “posterior distribution.” In LDA, the observed variables are the words and the hidden variables are the topic structure. So the point is to infer the hidden topic structure from the documents, called computing the posterior distribution (Blei). Thus, Catalinac uses a probabilistic topic model, the Latent Dirichlet Allocation approach, in her analysis of topics in Japanese electoral manifestos before and after the 1994 electoral reform.  She explores whether electoral reform makes candidates go from targeting the median voter (expected under Majoritarian systems) to groups of voters (expected with PR). Rather than specifying topics in advance, she uncovers the topics and estimates the probability that each document covers the topic. She then creates a term-document matrix: words in rows, document identifiers in columns, frequencies in cells. The word frequencies allow her to make inferences about the topics and to create estimates of probabilities that the topic will appear. She bases her assessment of particularistic versus programmatic topics on whether policies benefitted large or small groups (with particularistic obviously more geared to small groups.) She calculated the mean percent of discussions of particularistic and programmatic goods and compared the means before and after the reforms. She also used “wordfish,” which uses word frequencies to infer the locations of a document on an ideological scale (Slapin & Protesch).

We use supervised learning techniques because the differences between the documents we compare are preset and unambiguous. We do not need to scale or to sort the documents, as they have already been sorted into classifications. They constitute corpora of literature in different countries (the cross-national comparison) or works written at different points in time (the temporal differences), and our task is to compare the characteristics of these documents along dimensions that we have hypothesized from assumptions about the cultural differences that might give rise to diverse types of education systems. Because the differences among sets of documents to be compared are so rigid, we feel confident using word frequencies. Moreover, because we analyze literature instead of tweets or posts, there are likely to be a huge number of topics that have nothing to do with education.

We control for possible issues with the structure of the languages. It might be, for example, that a larger number of words are nouns in one language than the other, or that the tendency to form compound words in the non-English languages may alter the findings. We make a comparison of the English word “give” and the Danish “giver” to gain added confidence that the findings do not merely reflect linguistic differences. Moreover, the word frequencies vary in different directions across categories, and this gives added confidence that we are not simply observing linguistic differences.

**READERSHIP**

A set of questions about evaluating the impact of literature has to do with readership. At a minimum, the political class who makes public policy should have read literature, if fiction matters to perceptions of education reform. Yet to the extent that public opinion shapes policymaking in pre-democratic regimes, a broader swath of the middle class and even the working class should be exposed to these works for the argument to hold. Except for Britain, our corpora include all available full-text files and for Britain, we choose files that appear on online lists of works.

Fortunately, recent scholars of digital humanities have done remarkable work tracing the reading of texts, and there is ample evidence that fiction played a huge role in shaping the public’s perception of politics in the eighteenth and nineteenth centuries. Our own goals (and capacities) are more modest, because as political scientists, our focus is on the possible impacts of cultural artifacts on policy outcomes rather than on the patterns of use of cultural artifacts.

One author of this paper is conducting broader research that focuses on two types of readers. First, I explore how novels shaped intellectual elites’ views of politics and political reforms. Elite reading patterns give insight into how fiction influences policy decisions in authoritarian regimes, as literature was a site of political struggle and public opinion among elites. The literary community debated the grand issues in salons and taverns on a daily basis and this debate within the “republic of letters” led to social cohesion and shared values among the reading public. It facilitated debate among authors and spread their views to the broader reading public (Keen, 1999, 29-30). In short, Keen suggests that by the 1780s and 1790s, literature became the “single most effective means by which people could engage each other in a rational debate whose authority all governments would be compelled to recognize.” Thus literature was a kind of “group project” for projecting the groups’ interests onto the public consciousness so that “relations of power would give way to questions of morality.” (Keen 1999, 33). William Godwin reflected these views in 1793 when he wrote: “Few engines can be more powerful, and at the same time more salutary in their tendency, than literature. Without enquiring for the present into the cause of this phenomenon, it is sufficiently evident in fact, that the human mind is strongly infected with prejudice and mistake. The various opinions prevailing in different countries and among different classes of men upon the same subject, are almost innumerable; and yet of all these opinions only one can be true. Now the effectual way for extirpating these prejudices and mistakes seems to be literature” (cited in Keen, 1999, 28). Dissenters were particularly pronounced within this group (Keen, 1999 38). Romanticism at the end of the 18th century was a huge political project, and for Shelly, poets were de facto legislators (Keen 1999, 27).

Second, I explore how mass readers were exposed to the cultural assumptions of fictional works. For example, did these cultural assumptions play a role in how the masses thought about industrial institutions and democratic reforms in the nineteenth century? Of course, if the reading of classic works stayed within the elite class, literature may well have an impact on political reforms but be less influential on the development of the mentality of working class people.

In continental Europe, the revolution to read and educate in the vernacular – still a topic in early Danish literature in the 1700s – began in the sixteenth century with the Protestant reformation. Similarly, Henry VIII sought the printing of all English books to occur in London (Gilmont, 2003, 215-17). Reading was elevated in Elizabethan England, but the Protestant regime and civil war caused elites to fear literacy among the common man. The Glorious Restoration in 1688 was associated with diminished literacy. Yet the Society for Promoting Christian Knowledge was formed in 1699 to establish charity schools (Altick, 1954, 30-1). The reading public started to expand steadily in the early 18th century. Feather (1988, 90-91) suggests that 50 to 60 percent of men were functionally literate by the middle of the 1700s. Watt’s estimate is lower, calculating that newspaper readers were at most one in eleven in the mid-1700s, but that readership continued to expand to the end of the century. Two groups of people with more limited means were likely to be readers: apprentices and household servants (Watt, 36, 47). Mostly religious works read during the 18th century, but authors such as Defoe and Richardson crossed genres (Watt 50), and there is evidence that the middle and increasingly the working classes engaged with the leading fiction of the times. Keen believes that there was already a sizable working class readership by the late 18th century (Keen, 1999, 37). Certainly by the 1850s, the reading public fully encompassed a mass audience, and this was a period when there was a reading public of between 5 and 6 million people (Altick, 1954, 4-6.) Uncle Tom’s Cabin, the biggest literary phenomenon ever, sold 150,000 copies in the first six months (Altick, 1954, 6). Some believe that the Victorian writers were read by a more unified audience, compared with the fragmented readership for twentieth-century authors; Altick believes that publishers delivered cheaper fare to the working class, but some works such as Dickens’ *Household Words* was read very widely (Altick, 1954, 17-20). Indeed, in the mid-nineteenth-century, working-class autodidacts were a veritable movement and authors such as Dickens were like the rock stars of our times (Cordner). Just think of the dock workers, upon Dickens’ arrival home from America, shouting up, “What ever happens to Little Nell?”

The centrality of reading is reflected in book sales. It is beyond our scope to lay out book sales for all of the volumes in the corpora; however, the data for leading novels are instructive. Altick estimates the English population to be 6-7 million in 1750. Fielding published 6,500 copies of *Joseph Andrews* in thirteen months, and sold 5000 copies of *Amelia* in the first week (Altick, 1954, 49). The low sales in part reflected the high prices of books; however, this changed in 1774, when the publisher John Bell created a much cheaper venue on coarse paper that was the frontrunner of cheap reprinted six-penny volumes, and John Cooke followed suit with a series of British classics (Altick 1954, 54).

Robinson Crusoe was first written in 1719, before the huge growth of the book trade beginning around 1730, when first editions could total up to 10,000 copies (Feather 1988, 90-91). Yet eleven editions had been issued by 1759 (“Editions of Robinson Crusoe in English,” 1936, 22).

Readership of Robinson Crusoe expanded to encompass a broader cross-section of people when the novel was reprinted in the *Original London Post* (Watt, 42). In 1774, the publisher John Bell created a cheap book on coarse paper that was the frontrunner of cheap reprinted six-penny volumes, and John Cooke followed suit with a series of British classics (Altick 1954, 54).

The original issue of David Copperfield totaled a sale of 25,000 books, and 83,000 copies of the penny version of David Copperfield sold in three weeks in 1871. Dickens’s publishers had sold 4,239,000 works by 1882 in England alone (Atlick, 1957, 384.) Hardy’s *Jude the Obscure* sold 20,000 copies in the first three months (Altick, 1986, 238). Readership also expanded through the reprinting of cultural artifacts. For example, Ryan Cordell and David Smith have developed a website allowing readers to trace the circulation of “viral” texts through reprinting during the nineteenth-century. (See also Matthew Jockers).

We obtained a list of best selling novels for France during this period.

**DIFFERENCE OF PROPORTION RESULTS**

TABLE ONE DIFFERENCE OF PROPORTIONS RESULTS

(FREQUENCIES OF SKILL WORDS IN POVERTY SNIPPETS)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Period | Danish Percentage | British Percentage  | FrenchPercentage | Z Score | Significance level |
| 1740-1800 | 0.0619 | 0.0416 |  | 1.1876 | P < 0.235 |
| 1740-1800 | 0.0619 |  | 0.0047 | 3.252 | P < 0.001 |
| 1800-1860 | 0.0413 | 0.0289 |  | 2.423 | P < 0.015 |
| 1800-1860 | 0.0413 |  | 0.0062 | 4.4652 | P< 0.00001 |
| 1860-1900 | 0.0757 | 0.0233 |  | 5.693 | P< 0.00001 |
| 1860-1920 | 0.0757 |  | 0.0048 | 8.2588 | P < 0.00001 |

**DIFFERENCE OF PROPORTION RESULTS**

TABLE TWO DIFFERENCE OF PROPORTIONS RESULTS

(FREQUENCIES OF CHARITY WORDS IN POVERTY SNIPPETS)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Period | Danish Percentage | British Percentage  | FrenchPercentage | Z Score | Significance level |
| 1740-1800 | 0.0 | 0.079 |  | -4.0047 | P < 0.00001 |
| 1740-1800 | 0.0 |  | 0.0377 | -1.9953 | P = 0.046 |
| 1800-1860 | 0.0 | 0.068 |  | -10.1426 | P < 0.00001 |
| 1800-1860 | 0.0 |  | 0.0528 | -9.0095 | P < 0.00001 |
| 1860-1900 | 0.0072 | 0.037 |  | -6.1674 | P < 0.00001 |
| 1860-1920 | 0.0072 |  | 0.0691 | -9.8866 | P < 0.00001 |

**DIFFERENCE OF PROPORTION RESULTS**

TABLE THREE DIFFERENCE OF PROPORTIONS RESULTS

(FREQUENCIES OF SOCIETY WORDS IN POVERTY SNIPPETS)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Period | Danish Percentage | British Percentage  | FrenchPercentage | Z Score | Significance level |
| 1740-1800 | 0.6381 | 0.3078 |  | 3.9933 | P > 0.001 |
| 1740-1800 | 0.6381 |  | 0.6274 | -0.0852 | P = 0.9321 |
| 1800-1860 | 0.6656 | 0.4352 |  | 5.4746 | P > 0.001 |
| 1800-1860 | 0.6656 |  | 0.5266 | 2.877 | P < 0.004 |
| 1860-1900 | 0.5973 | 0.4175 |  | 6.9043 | P < 0.001 |
| 1860-1920 | 0.5973 |  | 0.5333 | 2.1064 | P = 0.035 |

**DIFFERENCE OF PROPORTION RESULTS**

TABLE FOUR DIFFERENCE OF PROPORTIONS RESULTS

(FREQUENCIES OF INDIVIDUALISM WORDS IN POVERTY SNIPPETS)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Period | Danish Percentage | British Percentage  | FrenchPercentage | Z Score | Significance level |
| 1740-1800 | 0.081 | 0.3161 |  | -4.9019 | P < 0.00001 |
| 1740-1800 | 0.081 |  | 0.1745 | -2.696 | P < 0.007 |
| 1800-1860 | 0.054 | 0.277 |  | -12.4826 | P < 0.00001 |
| 1800-1860 | 0.054 |  | 0.1476 | -5.924 | P < 0.00001 |
| 1860-1900 | 0.0904 | 0.2243 |  | -9.3919 | P < 0.00001 |
| 1860-1920 | 0.0904 |  | 0.1122 | -2.247 | P = 0.0247 |

**DIFFERENCE OF PROPORTION RESULTS**

TABLE FIVE DIFFERENCE OF PROPORTIONS RESULTS

(FREQUENCIES OF FAMILY WORDS IN POVERTY SNIPPETS)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Period | Danish Percentage | British Percentage  | FrenchPercentage | Z Score | Significance level |
| 1740-1800 | 0.2476 | 0.7737 |  | -5.7455 | P<0.00001 |
| 1740-1800 | 0.2476 |  | 0.7594 | -5.5435 | P<0.00001 |
| 1800-1860 | 0.1647 | 0.6441 |  | -12.506 | P<0.00001 |
| 1800-1860 | 0.1647 |  | 0.5996 | -13.1276 | P<0.00001 |
| 1860-1900 | 0.255 | 0.5835 |  | -9.1996 | P<0.00001 |
| 1860-1920 | 0.255 |  | 0.6839 | -12.771 | P<0.00001 |

**DIFFERENCE OF PROPORTION RESULTS**

TABLE SIX DIFFERENCE OF PROPORTIONS RESULTS

(FREQUENCIES OF STATE WORDS IN POVERTY SNIPPETS)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Period | Danish Percentage | British Percentage  | FrenchPercentage | Z Score | Significance level |
| 1740-1800 | 0.881 | 0.495 |  | 2.8564 | P < 0.004 |
| 1740-1800 | 0.881 |  | 0.2689 | 5.386 | P < 0.00001 |
| 1800-1860 | 0.7794 | 0.4929 |  | 3.111 | P < 0.0019 |
| 1800-1860 | 0.7794 |  | 0.4489 | 3.6318 | P < 0.0003 |
| 1860-1900 | 0.5054 | 0.4058 |  | 0.949 | P < 0.3426 |
| 1860-1920 | 0.5054 |  | 0.1928 | 8.2147 | P < 0.00001 |

**DIFFERENCE OF PROPORTION RESULTS**

TABLE SEVEN DIFFERENCE OF PROPORTIONS RESULTS

(FREQUENCIES OF RELIGION WORDS IN POVERTY SNIPPETS)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Period | Danish Percentage | British Percentage  | FrenchPercentage | Z Score | Significance level |
| 1740-1800 | 0.5143 | 0.2205 |  | 3.5276 | P < 0.0004 |
| 1740-1800 | 0.5143 |  | 0.2075 | 4.2294 | P < 0.00002 |
| 1800-1860 | 0.6853 | 0.2943 |  | 8.1033 | P < 0.00001 |
| 1800-1860 | 0.6853 |  | 0.4225 | 2.8644 | P < 0.004 |
| 1860-1900 | 0.4286 | 0.2126 |  | 7.4328 | P < 0.00001 |
| 1860-1920 | 0.4286 |  | 0.3405 | 2.861 | P < 0.004 |