MigratoryWords:

Computational Tools for Investigative Research

M.I.M.S Final Project
Spring 2010

UC Berkeley
School of Information

Jeremy Whitaker
Michael Manoochehri
Hyunwoo Park
Gopal Vaswani

Abstract

This Master's final project is the culminating effort of four graduate students at the University of California's, School of Information. It represents a semester of work exploring an interdisciplinary intersection of political, social, information, and computer sciences.

This project focuses on the applied process of investigative research as performed by journalists and researchers who scrutinize the domain of American public policy and political gamesmanship. Specifically, we focus on the role of political language and its ability to express flows of intellectual influence.

Aside from a general survey of the relevant interdisciplinary concerns, this paper discusses the role of open government at the intersection of private and public knowledge sharing. Our goal is to interrogate this intersection and to develop methods to identify influential language through natural language processing tools and customized search.

Our final product is an investigatory toolkit called MigratoryWords.com which allows journalists and researchers to query a massive index of text documents from multiple knowledge sources to get a broad sense of language's role in shaping the evolution of policy and the historical shifts in legislative discussion.

Table Of Contents

1. Problem Space

1.1. Introduction to the Problem Space

With recent progress toward more transparent government systems in the United States, new problems arise around how citizens can meaningfully engage and understand such a huge quantity of raw, and often context-less data. By extension, the new challenge isn't just finding the needle in the haystack, it's finding patterns that bind together individuals, organizations, information, data, and policy.

Meanwhile, a common public sentiment is that the political process skews disproportionately toward an overemphasis on polling trends and money, and citizens stress deep concerns of improper influence exerted by private industry. These concerns may be further validated by the recent Supreme Court ruling in Citizens United v. FEC, in which the court overturned precedents established by the Bipartisan Campaign Reform Act (BCRA) Section 203 and gave corporations and unions the right to use general treasury funds for "electioneering communications" and to advocate "for or against a candidate for federal office in the period leading up to an election" (Dorf, 2010).

While no single student effort could mitigate such foundational political challenges, we believe information systems research has a role in helping to develop tools which can increase investigative due diligence toward these issues. In particular, could equipping researchers and journalists with new tools help them to better study and report upon trends within the public policy domain? Specifically, can the transfer of unique language illustrate the sharing of knowledge between organizations? To this end, our project seeks to find novel methods of interaction with large bodies of political text and to experiment with methods that can highlight knowledge sharing channels that may not be overt.

Our hypothesis is that those who seek to influence will use rare sequences of terms, which when used together can suggest stronger emotional connotations. Commonly referred to as framing, the use of this constructed language becomes interesting when it occurs multiple times within a single corpora of publication. Even more interesting is when this constructed language migrates to new corpora as it breaks through organizational or medium channels of communication, reaching new audiences and perhaps suggesting successful propagation of one view to broader audiences. To put it simply, our strategy is to seek out rare combinations of words, which reoccur in multiple documents, across multiple sources of publication, and to track these over time, giving the user a sense of the temporal dynamics of language and knowledge transfer.

Our research goal will be to use computational methods to uncover these patterns within a huge cacophony of text. Will these patterns suggest trails of influence or draw a circle around group alliances? Is it possible to identify successful talking points that show a concentration of use during a short window of time? Are private organizations able to frame

topics in ways that are so effective that politicians pick them up as a means to push their legislative agenda? And is it possible to capture some glimpse of what might be taking place outside of the realm of public transparency?

1.2. A Case for Consideration: The Marketplace of Ideas

Published on November 14, 2009, New York Times journalist Robert Pear wrote a story focused on a collection of e-mail messages obtained by the newspaper. The e-mails were produced by a lobbying firm under the employ of biotechnology company, "Genentech and...two Washington law firms" (Pear, 2009). The contents of these e-mails were authored to assist, persuade, or influence both Republican and Democratic leadership. While such a scenario is common by today's political standards, what may be more remarkable is that some 42 members of Congress submitted congressional statements using text derived from these documents, either verbatim or with minimal editing ("LouisDB.org", n.d.).

To track this transfer of information, Pear applied his professional intuition to the text to determine which language blocks were most interesting. With these blocks he performed iterative manual searches against the Library of Congress database, a system of public records that contains, among other corpora, the Congressional Record. Within the results were a collection "revise and extend" remarks, placed by members of Congress, as footnotes to the *Affordable Health Care for America Act* of 2009("LouisDB", n.d).

As described by Pear, the objective of the e-mails was not to stop or even modify the legislation, rather the authors sought to explain complex topics in more understandable terms while framing the fundamentals of the conversation in ways favorable to Genentech's business interests. In particular the efforts sought to shine favorable light on Genentech's ingenuity and to suggest that new "biosimilar" variations of preexisting Genentech patents should be protected in the international marketplace. The results were accomplished via two different arguments and delivered by two different emails, each one tailored to the agendas of the two dominant political parties:

> *Democrats emphasized the bill's potential to create jobs in health care, health information technology and clinical research on new drugs.*
>
> *Republicans opposed the bill, but praised a provision (favorable to Genentech) that would give the Food and Drug Administration the authority to approve generic versions of expensive biotechnology drugs, along the lines favored by brand-name companies like Genentech.*
>
> *(Pear, 2009)*

In effect, Pear had discovered a concentrated campaign as lobbyists pursued what they internally described as "aggressive outreach" to "secure as many House R(epubulican)s and D(emocrat)s to offer this/these statements for the (Congressional) record as humanly possible" (Pear, 2009).

While this case may seem troubling on the surface, neither Pear nor the persons he interviewed saw this language coordination as a nefarious activity. In fact, sources quoted in Pear's articles emphasize that the knowledge sharing activities of Washington are normal and can be beneficial to the process of getting to the best policies. If one removes the enveloping financial implications that accompany this corporation's involvement with politics, most people would not find fault with a member of Congress turning toward domain experts to help parse the complex fields of bioscience or pharmacology. In fact, a politician should be encouraged to apply her best judgment using any available resources, as long as the decisions process is ethical and result in policy that is well grounded and best serves her constituents.

From this perspective, the efforts of lobbyists and think tanks are often seen as a welcome contribution by those in policy circles. And while one can certainly find signs of an "age-old tension between the world of ideas and the world of policy" (McGann, 2007, p. 7), this tension delivers bi-directional benefits that can serve the citizenry through improved foundations of policy. The efforts of campaign finance reform and of Citizens United v. FEC are a matter of finding the right balance between citizen's rights, corporate rights, free speech, and economics. They are not about eliminating influence, but rather shaping it to yield the best compromise and most reasonable results.

Without making judgment on whether these shifting scales constitute good governance, we believe Pear's scenario describes an interesting social, political, and computational problem. Specifically, can an information system help a journalist to identify similar patterns without explicitly knowing where to begin? And could the patterns provided inform deeper research as the journalist applies their traditional tools and professional intuition along an extended line of enquiry?

2. The Intersection of Political and Social Science

2.1. Social Science Considerations

Before applying a technical approach or delving into project specific considerations, it is valuable to consider the complexity of influence as it is understood by different academic domains.

As a transitive verb, Merriam-Webster (2010) defines the term INFLUENCE as follows:

i. to affect or alter by indirect or intangible means
ii. to have an effect on the condition or development of

The political realm offers a more directed definition: the "ability to get others to act, think, or feel as one intends" (Banfield, 1962, p. 3).

For sociology and social psychology, there are different assessments for social influence and political influence, derived from views on power and capital. All are relevant to our research as political actors regularly engage in complex exchanges of reputation, resources (implicit and explicit), trust, and obligation.

Political influence is rooted in Max Weber's theories on power and the state, where state-based power asserts a downward pressure (or influence) on citizens. This power is derived from the state's monopoly of the legitimate use of physical force (Weber, 1918). Of parallel interest to this project is Weber's identification of those who exist at the periphery of policy making who seek to "engage in politics...to influence the distribution of power within and between political structures" (Weber, 1918).

In social influence theory, influence flows between actors in a network where the roles of individuals are dynamically shifting dependent on actor attributes in relation to other members of the network (Friedkin, 1999, p. 2). Within the interactions of politics and organizational theory, this category of influence can be seen where social influence and reputation help to reduce social impediments to successful knowledge transfer. (Disterer, 2001; Lucas, 2005, p. 88).

As identified by Bourdieu and Wacquant (1992), the root of this transfer is "capital", which "take(s) a variety of forms (and) is indispensable to explain(ing) the structure and dynamics of differentiated societies" (Miller, 1992, p. 119). These forms of capital apply themselves differently to political transactions, first as traditional social capital in "durable networks of...institutionalized relationships of mutual acquaintance and recognition" (p. 119).

In parallel to social capital are flows of cultural capital or "informational capital" (p. 119) which take "embodied, objectified, and institutionalized" forms. One can view these forms of capital in the knowledge transfer process when considering how reputed intellectual leaders such as top research schools or influential domain experts are able to brand their intellectual products with high efficiency and to capture wide audiences by using embodied or institutionalized capital as a mechanism to minimize competition.

Finally, Bourdieu posits that "social capital...constitute(s) a kind of political capital which has the capacity to yield considerable profits and privileges, in a manner similar to economic capital in other social fields". For us, the key point is that Bourdieu delineates between social forms of transacted capital, and economic or financial capital, arguing that both can offer a similar result. In context consider how much easier it is for us to measure the explicit data of well expressed financial transactions, but how challenging it might be to definitively state the implicit agenda of those exchanging financial capital for a service. Does a campaign contribution always mean what we think it means? (Lessig, 2009) While humans *may* be able to make informed assumptions using only this financial information, algorithms are certainly poorly suited to this task.

Weighing these considerations, we believe another mechanism for analyzing influence might come by understanding how the transfer of "informational capital" can be measured. Specifically, can we identify the informations flows of those who are seeking to gain *intellectually derived influence*; that is the transference of powerful ideas or expert knowledge based on reputation or correctness. It is our belief that this is the actual asset which financial influence is seeking to purchase, and it is this transfer of intellectual influence which we would most like to capture using our toolset.

2.2. The Political Application of Influence

In looking for intellectual influence, our project searches for language patterns that express some informal and amorphous relationships. Successful identification of these patterns however cannot be taken as declarations about intent, but rather they highlight interesting patterns that are worthy of further human research. Results should not be taken to imply causation, where one action leads to specific reaction. Instead, we expect our users to apply due diligence when coming to their own conclusions.

We take this stance because we recognize that political decision making is a very complex process and that the behavior of each actor depends on independent "weighing or aggregation" (Banfield, 1962, p. 327) of the complexities emphasized in sociology's theories on power and social influence. No political decision can be made in a vacuum where only intellectual and financial factors effect the outcome. Social factors must be considered.

A Pocket Theory for Social Factors of Political Influence

One undeniable force in these social theories is how reputation and social networks drive transference. Some modern thinkers have applied physical metaphors of movement to describe the dynamics involved in this process.

> *Information cascade occurs when individuals, having observed the actions and possibly payoffs of those ahead of them, take the same action regardless of their own information signals (Bikhchandani, 2005, p. 1).*

By using the term cascade, Bikhchandani and others (reputational cascade (Sunstein, 2007, p. 85), tipping point (Gladwell, 2000, p. 7)) illustrate how information flows can become exponential. In turn we apply similar metaphors to our investigation of how expert knowledge can transfer with greater efficacy when under certain network effects, creating what we refer to as an influence cascade.

Narrowing this analysis to the political domain, the *influence cascade* involves a set of modeled actors, each with a different set of traits. It supposes a flow of information from researchers and domain experts, to other levels of the political process. The flow of information can suggest a kind of momentum as it travels from a narrow point of origination, gaining energy as high-reputation organizations and individuals buy-in to the policy at different stages, moving progressively outward toward a larger audience. Of course, this is not a one-way flow and failed campaigns can offer feedback for domain experts who can choose to reframe their argumentation or to seek out better channels for influence distribution.

In this influence cascade, the following base assumptions are made:

- There is an information asymmetry and/or a time asymmetry as individual politicians seek out expert knowledge and guidance to help them grapple with complex topics.
- Some organizations specialize in providing expert knowledge, these include: think-tanks, lobbying organizations, independent experts, organized citizen groups, non-profits/NGOs, private companies, and academic institutions. These organizations promote policy alternatives by:
    i. offering nonpartisan policy suggestions
    ii. pushing agenda based political guidance
    iii. unconditionally generating open-source expert knowledge (i.e. Academia, NGO shared expertise)
- Political leadership within network groups wants to harmonize political intention across their spheres of influence. They aggregate political knowledge as central to their political platform and use intra-network reputation and trust to build momentum. Either internally or with expert assistance, they frame certain political language to be highly consumable to mainstream audiences.
- Momentum of key individuals or a collection of several actors in support of the issue helps to create the start of the cascade which flows outward toward media. Media organizations identify the theme and repeat it until new framing appears or the topic loses momentum.

The behavior of the influence cascade is as follows:

1. Expert knowledge is propelled through a network of actors via reputational transfer along channels of minimal resistance.

2. Momentum in the cascade is sustained at each node by actors who lack their own personal expert knowledge and/or who suffer from their own information/time asymmetries.
3. In opposition to the flow, nodes who have differing expert knowledge personally, or who apply another view with support from alternate sources reputational/expert knowledge, can alter momentum by creating resistance through disputation.

3. The Intersection of Political Policy and Information Policy

3.1. Assessing the Political Status Quo

To understand how outside groups might seek to influence policy at the Federal level, it is useful to briefly survey the types of private organizations most directly embedded in the process of policy formation. For our purposes, we emphasize think tanks and lobbying organizations as they both serve to aggregate many different inputs and often encapsulate the influence efforts of non-profits, academia, corporations, political-action-committees, and other government entities. These two institutional types are expected to apply very different methods of asserting their influence, but opinions differ on the truths of the matter.

a. A Generalized Landscape View of Think Tank Influence

There are 1777 think tank organizations in the United States. Obtained from IRS documents in 2004, the combined operating revenue of twenty of the largest think tanks was approximately $631 million USD (McGann, 2007, p.23).

As outlined by James McGann, a former PEW Center employee and the proctor of the University of Philadelphia's ongoing international survey of this industry, think tanks are primarily research institutions who are organized to help with the development of informed policy decision making. They are registered as not-for-profit institutions and are prohibited "from attempting to influence a specific piece of legislation (and they generally) understate rather than overstate their influence on major policy issues" (McGann, 2007, p. 3).

Think tanks often "act as a bridge between academic and policy-making communities, serving the public interest as an independent voice that translates applied and basic research into a language that is...accessible to policymakers and the public."(p. 7) In the views of one think tank staffer, "rarely does an idea leap from a think tank to become public policy. More often ideas contribute to the national debate and influence the political climate in indirect ways. Sometimes that influence can be substantial." (p. 4)

Irrespective of the expectations of impartiality that accompany their non-profit status, as the industry has expanded, "the role of many...organizations shifted from providing objective, scholarly research...to disseminating...action oriented policy that aimed to influence the decision-making process."(p. 2) Of particular interest are groups which McGann identifies as advocacy think tanks (ex. CATO), policy enterprise organizations (ex. Heritage), or

party-affiliated think tanks (ex. NBER). These groups, while academic in some regard, bias their policy suggestions toward organizational objectives.

Despite efforts to construct methods to track the influence of think tanks, "determining the extent to which a think tank or group of think tanks influence(s) a particular policy decision remains a daunting methodological task." (Weidenbaum, 2009, p. 90) Instead the sector relies on "anecdotal evidence of...impact indicators" such as policy adoption, policy change, and policy implementation (McGann, p. 42). Considering the assessments of these industry experts, we believe that a computational approach could be applied to this problem space by using aggregate scoring of a number of different values, including linguistic statistics.

b. A Generalized Landscape View of Lobbyist Influence

As of 2009, there were 13,694 registered federal lobbyists in the United States. The total combined spending on lobbying activities was $3.46 billion USD. ("Lobbying Database", 2010)

Lobbying organizations take roles in issue advocacy on the behalf of corporations, public/private interest groups, foreign governments, or even other federal agencies while protected by First Amendment provisions on the right to assemble and petition the government. As private organizations they are generally hired for their services, but in some cases they work pro bono for non-profit organizations.

While often motivated by economic incentives, lobbying organizations must follow regulations that guide and track their behavior, including a host of federal campaign finances laws such as the Lobbying Disclosure Act of 1995.

The basic provisions of Lobbying Disclosure Act require all lobbyists to be registered with the government, to disclose their employer, and to declare the issue areas of focus. With specific regard to transparency, Congress declares in Sec 2 U.S.C 1601 that full disclosure is essential to increasing public confidence in the government process with regard to influence and accountability. Failure to comply with these regulations can incur criminal and civil penalties of up to $200,000 and five years imprisonment.

Sec 2 U.S.C 1601

> *The Congress finds that—*
>
> *(1) responsible representative Government requires public awareness of the efforts of paid lobbyists to influence the public decision-making process in both the legislative and executive branches of the Federal Government;*
>
> *(2) existing lobbying disclosure statutes have been ineffective because of unclear statutory language, weak administrative and enforcement provisions, and an absence of clear guidance as to who is required to register and what they are required to disclose; and*

> *(3) the effective public disclosure of the identity and extent of the efforts of paid lobbyists to influence Federal officials in the conduct of Government actions will increase public confidence in the integrity of Government.* *(cite:Senate.gov, 2010)*

While lobbying is widely disparaged by portions of the American public, a 2010 Pew Center poll (completed one month prior to the Supreme Courts ruling on Citizens United v. FEC) revealed that, in light of then current economic and political considerations, further reforms to control lobbying activities were not a public priority ("Public's Priorities for 2010", 2010). Although, in the wake of the Citizens United decision, there has been significant response from many interest groups and political leaders. On April 29, 2010, Senators Schumer and Holden introduced the *DISCLOSE Act* which aims to plug the gaps created by Citizens United, and President Obama spoke directly on this issue stating "powerful special interests and their lobbyists should not be able to drown out the voices of the American people" ("Statement by the President", 2010).

Irregardless, as previously mentioned, from the perspective of many insiders, disclosed lobbying plays a vital role in the process of percolating ideas into the political discussion space. While applying greater financial capital toward lobbying a cause can exert greater pressure through increases in quality of services or via strength in numbers, the financial capital itself is not the enemy of progress. If one assumes that lobbyists target politicians who have political self-interest in aligning with their cause, then ultimately, properly equipped voters should be empowered to determine whether the politicians have acted appropriately for their constituents. However, problems may arise if voters have an information asymmetry about what motivates the decisions of their political leadership, preventing them from properly exerting checks and balances on the process. While the Supreme Court's 5-4 majority currently believes transparency measures are sufficient there are many who are not so certain. In this scenario, full disclosure and transparency are the mechanism to measure and maintain accountability, but the laws were never designed to prevent lobbying organizations from applying influential pressure on the generation and propagation of new ideas, they are simply designed to apply regulatory pressure on the companies who fund the actions of lobbyists, and the politicians who fear that corporate influence may run counter to the will of their constituents.

c. A Better Lens for Assessing Political Influence?

In the wake of Citizens United and as favorability ratings for Congress reach record lows (Pew, 2010), increasing accountability and voter trust are seen as critical to the health of American democracy. One could make a strong argument for even greater transparency into how policies are coordinated between private and public interests. However, raw and context-less financial indicators while an easy means of tracking the obvious, generate vitriolic reactions about corruption and make it very difficult to identify the causative influences of these organizations on policy making.

At the core, we believe Americans aren't overly concerned with minutiae, but rather are most interested in knowing the sum of the game. Are public policy organizations really motivated by the best ideas? Are politicians recognizing good policy and acting in the best interests of our citizens; or are they overwhelmed with political gamesmanship instead of

focusing on governance? Most citizens, if asked these questions would respond with great skepticism, and while this skepticism may be justified, perhaps what is needed is a better lens by which to view the process. If citizens knew that 70% of the policy coming from the Government was based on sound advice and in the public's best interest, would they have greater patience with the process of governance?

3.2. Open Data: Concentrated Political Speech or Overflowing Quagmire

To best understand the intersection where private influencers and public policy meet to form legislative output, it's useful to consider how the government delivers policy information to its citizens, and to consider how open data can best serve the efforts of citizens to engage and create context from complex datasets.

Advocates for Open Government view the internet as a "primary source of information" ("Improving Access to Government", 2009) and see it as a potential "fifth estate" ("Improving Access to Government", 2009) provisioning a new set of stakeholders with their own mechanisms to check and balance government activity through increased citizen involvement and connectedness with the political, social, and informational capital inside the system. Similarly, governments also recognize the benefits brought through greater public involvement as it becomes easier to facilitate issue cohesion via online campaigns, to increase fundraising via social networking, or to build political capital by advocating increased public scrutiny (into areas deemed "safe"). While these efforts bring alignment with the Brandeisian value that "sunlight is the best disinfectant" there is a trade-off between government autonomy and public trust. Therefore, different politicians have highly different views on the matter.

The roots of transparency extend to the early days of American democracy when the founders helped to establish civic collectivism by maintaining open and publicly documented legislative proceedings. These records exist in one form or another all the way back to the Continental Congress of 1774. Now managed by the Government Printing Office and the Library of Congress, the legislative bodies generate a huge outpouring of text annually. In 2009, the combined total for legislative documents published by the Library of Congress was 71933 discrete URIs (Appendix 1), an average 197 new URI instances per day.[1]

Continuing this tradition, on his first full day in office, President Obama issued a *Memorandum on Transparency and Open Government*(Obama, 2009) which stated that his administration would seek to provide "unprecedented" access to government information, and to stimulate citizen participation and collaboration via online tools. Building on this initial declaration, the administration developed the *Open Government Initiative* and *Open Government Directive* which commissioned the launch of the data.gov portal and a

1 The contents of each URI in this corpora is highly variable spanning from a single page to up to 99 pages per URI.

host of affiliated services. Data.gov has scaled dramatically between January 2009 (333 datasets) and May 2010 (1780 datasets) ("Growth of the number", n.d.). It was judged by independent groups to be the "strongest and most comprehensive lobbying, ethics, and transparency rules and policies ever established by an Administration" ("Democracy 21", n.d.).

As noted by Robinson et al. (2008), government data systems are currently lacking advanced features that are tailored to types of use. Instead the government approach is to publish en-masse and applies little consideration to "how" people wish to engage with the data provided. By attempting to organize and distribute both presentational and infrastructural aspects of this information distribution, and doing neither particularly well the authors suggest that government focus all efforts on the infrastructure, providing open and standardized architectures, and distributing content in highly-consumable open formats. This approach allows government to focus on data security and privacy, while balancing extensive regulatory and compliance challenges (Robinson, Yu, Zeller, & Felten, 2008, p.11). This concentration allows the government to ignore the creation of task-focused presentational systems equipping citizen and interest groups with tools to develop platforms tailored to the specific use cases determined by their user communities.

Instead, with the government efforts spread too thin, current open data systems are often incomplete and do not follow best practices for re-use. In response, non-profit advocacy groups such as The Sunlight Foundation and The Center for Responsive Politics, commit resources to cleaning and marking up government content. With better structured data, these organizations channel communities of users toward these new data sets, where citizens can build custom applications designed to provide innovative new uses that better serve shifting public needs. For example, in 2009 the AppsForDemocracy contest was able to stimulate construction of 47 mobile apps in 30 days, with only $50,000 in seed capital. The results were valued at $2.3 million and praised by Vivek Kundra (former D.C. CTO and current Obama CIO) as "produc(ing) more savings for the D.C. government than any other initiative." ("Apps for Democracy", n.d.). Similarly other campaigns like The Sunlight Foundation's Apps for America contest connect funders like Adobe and Google with groups of citizen programmers who seek to address challenges designed by Sunlight to serve their goals for government accountability through transparency.

3.3. The Implications of Greater Transparency

One justification for the ruling in Citizens United v. FEC was the courts' assertion that current accountability and disclosure measures were sufficient in themselves to hold corporations accountable. If this is the case, then the role of transparency projects is all the more critical. However, in his dissenting opinion Justice Stevens added that while "modern technology may help make it easier to track corporate activity, including electoral advocacy...it is utopian to believe that it solves the problem" (Citizens United, n.d.). The majority has stated that financial disclosures justify full and unrestrained corporate speech, but is this not ignoring other avenues of influence that might be more difficult to

track? How are we to know when a politician is promoting corporate interest at the detriment of his constituency? With current systems how can we see when corporate agendas become political speech, and by extension, become political law? It seems critical that the courts support the constitutional interests of the constituency and their role as the electorate, and we believe current practices are insufficient for them to detect all measure of influence.

One example mentioned previously, is the building of conclusions from raw financial data without appropriate context or sufficient justification. Lawrence Lessig notes that "even if we had all the data in the world and a month of Google coders, we could not begin to sort corrupting contributions from innocent contributions". Instead we should be cautious of "where and when transparency works, and where it may lead to confusion, or to worse" (Lessig, 2009).

Similarly, raw transparency alone doesn't offer a panacea and increased data availability to wider audiences carries broad implications and justifying an informed and cautious approach. In particular, when the contents of datasets reveal personal information or when inaccurate correlation too easily leads users to false causation. Not every dataset should be exposed, and not every conceivable reuse of data will yield socially positive results.

With these considerations in mind our project seeks to apply "targeted transparency" and to explicitly focus on intellectual influence. We aim to put tools in the service of human interpreters who are well qualified to apply their professional filters and intuition to determine how to best direct further inquiry. If there is an ethically solid justification for a journalist to extend their inquiry by layering financial data upon the language patterns revealed by our tools, then we certainly encourage this, however we can not be certain that a computational merging of these datasets can provide accurate correlations in all cases.

3.4. Policy Disclaimers: Privacy and Fair Use

On Privacy

While the rights of individual privacy should be a parallel consideration to anyone who advocates for greater transparency, we do not feel that we are creating new dangers for individual or institutional privacy. As discussed more fully in section 4.2, our specific interest is with public domain data that has been released for public consumption. These data sets are an accounting of the activities of publicly elected officials and their actions on behalf of constituents and with this in mind, all parties are fully cognizant of their responsibilities and required accountability.

As mentioned previously we recognize that with projects that depend on large sets of statistics and data, there should be concerns over implications of inaccurate correlations between different corpora. This is especially true where application programming interfaces (APIs) allow users to remix dataset in unintended methods, not originally intended by the developers. We hope that all uses of our API interfaces are done with respect to personal privacy and apply common ethical standards.

On Fair Use

"However firmly liberty may be established in any country, it cannot long subsist if the channels of information be stopped,"

Senator Elbridge Gerry (1792)

For our project to be successful, we must interrogate both public sector and private industry content. However, all content which we reference is publicly available by the rights-holders at front-facing points of distribution. Acquisition of their non-standardized, non-aggregated content was time consuming, but critical in order to measure any forces which might apply influence at the intersection of private and public domains.

In a manner similar to Google ("Google Cache", n.d.), we apply fair-use doctrine to our caching of external content from non-governmental websites. In all cases, we show five full sentences of text for contextual perspective, and whenever possible, we provide clear links returning to the initial source publisher.

We performed our caching in accordance to the rules of each domains' robots.txt ("Web Robots Pages", n.d.) files and we also provide clear steps for rights-holders to have their content removed from our system. We do these actions in good-faith, as there is no clear precedent that robots.txt files are cause enough to invalidate fair use ("Web Robots Pages", n.d.).

Additionally, in accordance with 7 *USC* § *107* ("United States Code", n.d.), our intentions explicitly focus on research and investigative purposes. We have no for-profit intentions with this service and we view our usage of all data as transformative adding new mechanisms for of interpretation via timeline and toolset functionality which empower users to do research and exploration of data in new ways.

Most importantly, this project has sought out this rich dataset in order to to bring new context to existing government and legislative documents and to enrich the potential experience for citizen users. We believe that it is important that overly strong intellectual property restrictions not detract from the ability of citizens to gain knowledge about public behavior; and that projects which bring richer understanding of the involvement of private interests in the public political process are especially valuable to a healthy debate.

4. The Intersection of Disciplines on this Project

4.1. General Approach

To track influence and discuss the coordination of ideas, our project collected a large index of publicly accessible text from non-governmental policy, media, and corporate organizations (discussed more extensively in Sections 4.2 and 6.3). We focus on these channels as they represent a concentrated source of attempts to influence trends in public opinion and policy. Starting with the non-governmental groups, we compare their text to governmental records in hopes of finding shared "interesting language" – that is, constructed language that rarely occurs without explicit messaging intent, but which is also significant in that it is shared across multiple document instances.

Initially our sources exist as separate entities, but our intention was to find connecting language similarities across corpora and so we applied a content structuring and reorganization process to merge and operate against multiple datasets. Finally we applied natural language processing techniques (NLP), and augmented our discovery process by using available metadata to add further context.

In summary, the result we produced is a toolkit for researchers and journalists that takes available public document resources, identifies unique and interesting language, and collocates the occurrences of this language across multiple documents, published by multiple organizations, in multiple corpora. To accomplish this result we perform a term frequency analysis, sorting with preference to rarity. We cover further technical methods at length in Section 6.

Our goal has been that this tool should identify the following:

i. What document in the index represents the initiation of this language?
ii. What documents in the index rely on the same language?
iii. Who are the organizations who use the same language?
iv. What is the overall context of this language across the discussion?

4.2. Content Sources

In accordance with provisions of fair use, we began to assess language influence by collecting and structuring information provided by the following sources:

Sources of Media, Policy, and Corporate Speech

Press Release Documents (PR)

Corporate and media press releases via content provider, PRNewswire, published between September 2009 and March 2010, and the entire span of years 2006 and 2007. This source contains 510,018 records and 4.1gb of text data.

PR Documents are produced by a wide cross-section of organizations, from corporations, non-profits, state and federal government branches, universities, and others. Designed to channel messaging toward stakeholders and to alert outsiders of developments. PR documents represent a broad cross-section of constructed speech designed to channel information to the highest volume possible.

Think Tanks (TT)

With over 5,500 public policy think tanks worldwide, we needed to place a reasonable limit on the volume of our index. Based on the prominence of recent political debate with regard to health care during late 2009, we identified the top 11 Think Tanks in the domain of Health Care using the University of Philadelphia's "Global Go-To Guide", which offers an extensive survey of the policy advocacy community (2009). From these organizations, we acquired every available position paper spanning all topics (not just health care). These sources contain 66,049 records and 2.4gb of text data.

1) Brookings Institution	7) Cato Institute
2) National Bureau of Economic Research (NBER)	8) Fraser Institute
3) RAND Corporation	9) Center for Global Development
4) Urban Institute	T10) Civitas
5) American Enterprise Institute – USA	T10) National Center for Policy Analysis (NCPA)
6) Council on Foreign Relations Global Health Program (CFR)	

As non-profits designed to coordinate the reformulation of academic ideas into consumable policy, the think tank sources offers a means of viewing public policy gestation and propagation. This set of think tanks, influential in the space of health care, offers a cross-section of this community by providing an interesting set of large generalist organizations and small specialist organizations.

Corporate Position Papers (CPP)

Unable to find a large repository of lobbying white papers, we used a reasonably sized collection of partisan position papers aggregated by Washington D.C based newspaper, The Hill ("White Papers", n.d.). In total, this index contained papers published by 107 organizations in 83 area topics totaling 500 records and 29mb of text data.

Another collection of constructed text, this group of documents represents organized efforts by corporations or lobbyists to educate or influence. Our assumption is that lobbying organizations are hired to take these campaigns directly to policy makers and we are interested to see if we can trace a campaign's propagation. [2]

Source of Government and Legislative Speech

Government and Legislative Speech (GR)

Government documents acquired via the Sunlight Foundations's LOUISdb (http://louisdb.org) archive. Data spans from late 1999 to early 2010 and includes sections titled: Bills and Resolutions, Congressional Reports, Congressional Record, Congressional Hearings, Federal Register, and Presidential Documents. This sources contains 584,326 records and 11.4gb of text data.(Appendix 1)

As a deeply ingrained and formalized system of open governance, government documents represent a point of consolidation where tested policy designs become widely publicized as part of the legislative process. Not limited to successful legislation, even failed attempts at policy are captured, while revise and extend formalities allow majority and minority congresspersons to declare their specific views, either individually formed or constructed along party lines. In parallel, congressional hearings represent the gathering of valuable outside perspectives that are relevant to informing the decision-making process. As a complete public record these documents are widely available, but often difficult to parse for the individual citizen.[3]

4.3. Potential Social Benefits

> *"Increasing the ability of the public to discover, understand, and use the vast stores of government data increases government accountability and unlocks additional economic and social value." (ConOpsFinal, 2009, 6)*

Early on in our project designs we identified our potential user base as "the muckrakers": those who dug tirelessly to uncover important information in an effort to protect those who were not able to defend themselves against larger and more powerful forces. Ultimately these individuals are knowledge detectives and any tool that shortens the path toward increased justice was a good start.

In the midst the current crisis in media there is an ongoing reassessment of priorities as we seek to define what was good about "old" journalism and what should live on in "new" journalism. For those involved in this reassessment, what is nearly universal, and likely the

---

2 This corpora represents one of the smallest in our index as we were unable to uncover other major aggregators of lobbyist text. Given more time there would be utility in caching more of this text, but another major effort of data acquisition would be required to design custom scripts to gather data from individual sites in a manner similar to how we cached think tank organizations.

3 One factor of concern with this corpora was the level of detectable noise from excessive salutations and other non-legislative material. However, our results seem to mostly avoid this by using multiple occurrences across multiple corpora as a filter. See Section 3.5.

single most widely accepted truth, is that investigative journalism *must* live on. The success of organizations like ProPublica, the Hellman funded Bay Area News Project, and the rapid expansion of the Center for Investigative reporting are proof that interested benefactors are responding to dramatically shrinking investigative news budgets across the country by finding alternative means of supporting this vital contribution.

For information systems experts who recognize this crisis and feel similarly about the importance of journalism's value to a healthy democracy, it is vital for us accept some responsibility in shaping new technical approaches that can help to strengthen journalism and to support its most essential forms through this metamorphosis.

Also valuable to us as system designers, journalists may present an ideal type of user; one with well honed skills of inquiry, an active commitment to ongoing political engagement, and a host of non-computational skills that can act in parallel to our computational systems. When combined into an investigative team, the skilled journalist and the task focused computational system can achieve greater ends. Take for example a recent story in Wired magazine discussing the abilities of average human chess players who competed against supercomputers and grandmasters of the game. These average humans, augmented in their skills by computer systems, dramatically outperformed both computational and human experts. The key was the user being "especially skilled at leveraging the computer's assistance" (Thompson, 2010, p. 42).

Concurrently, journalists should know best about when the application of human sensibilities outweighs the need for "naked transparency" (Lessig, 2009) and when careful disclosure practices should be applied to big social and political issues. While it is these big social and political issues which make transparency meaningful, a good journalists can offer context and nuanced perspective that help readers to get complete information, rather than simply providing hard, cold facts that can skew toward emotional, unbalanced interpretation.

Peripheral to assisting with journalistic efforts, its is easy to see how these augmented search tools could be used by linguistic, political science, and social science researchers who are trying to track language behavior.

Over the long term, tools like this, used effectively by journalists and researchers, might lead to a better informed public which applies more scrutiny to the language used in political campaigns and commercials, governmental legislation, and corporate speech. Under ideal circumstances, being able to see widespread influence factors may help citizens to have a more informed view of the roles of influencers who exert pressure on the political system. In some cases the result might be increased accountability measures, in other cases, perhaps highly coordinated campaigns between public policy organizations and politicians can be better represented as serving the public's best interest and highlighting how intelligent solutions can be devised through open innovation and collaboration.

5. Collaborative Research Process

5.1. Qualitative Research with Potential Users

Needs Assessment with Investigative Journalists

To inform our research we began our inquiry with meetings and interviews with several prominent journalists and investigative researchers.

In these meetings our goal was to better understand the investigative process and to consider how these specialists interact with large public data sets as they develop a story. How did they approach an investigative effort? How has increased accessibility to government data changed their approach to story building? Did they use commonly available mainstream search tools or more advanced tools designed specifically for their query?

What we learned was that the story gestation process is highly iterative and involves adhoc ways of traversing human connections and statistical resources, which included aggregating different sources, playing details against each other, identifying difficult patterns, and triggering input from trusted authorities (either human or institutional). A resource in this of course is understanding what is interesting to readers and applying human intuition as the investigator, equipped with strong supporting details, leads the reader through obscure or less interesting minutiae to the more compelling storyline. These human factors are difficult to reproduce and there was some skepticism with regard to how some current computational efforts over promote "low effort" shortest path linkages between datasets as conclusive without applying human diligence.

However, computing tools are seen as expanding the resource toolkit for these professionals. Each spoke of specific tools that could assist with their line of questioning, but the primary dependencies were on general "search and repeat" approaches. A few different investigative journalists spoke of new tools which allowed for correlation of specific known public facts such as census or public disclosure information (property values, economic indicators, etc). More technically advanced journalists identified an interest in custom scripts or data enrichment efforts via correlation. Another investigative journalist identified an interest in understanding how political language gets re-used in California politics and stated an ongoing desire for a tool that helped identify "congressional remixing" of prior (failed) legislative efforts into new attempts.

Social Science Discussions with Academics

In addition to these journalists, we spoke with two UC Berkeley professors who specialize in language and information transfer. They provided feedback about how our project could be framed within the fields of politics and sociology. One was particularly interested in understanding how our project may fit into the John Dewey-Walter Lippmann debates on

how, citizens can be informed participants in the democratic process amidst a flood of political opinions and information. While Lippmann argued that a class of governing citizens and elites were necessary to maintain democracy and navigate the complexities of modern government, Dewey felt that common citizens should be empowered by better access to our increasing amount of political information. Furthermore, one of the professors specifically emphasized an interest in delineating how our computational system might differ from the methods humans currently use to accomplish the same task - is our goal to replace or augment the role of a human researcher? Or does our project attempt to tackle a problem space that has yet to be explored without computational techniques.

Policy and Needs Discussion with Legal Staff at Non-Profit Advocacy Center

Finally, we interviewed a policy counsel for a non-profit open data advocacy group who was interested in our computational approach to linguistic influence. In particular, could we help in the visualizing how "omnibus" bills are recycled from previous legislation? Also, could linguistic analysis help to identify similarities between public testimony of registered lobbyists and text that appeared in the Congressional Record.

Collaborative discussions with stakeholders

In parallel, the team held regular meetings with a fantastic group of UC Berkeley faculty and outside collaborators. These participants included:

- Rob Ennals, Research Scientist with Intel Research Berkeley focusing on problems of misinformation and bias on the web.
- Laurent El Ghaoui, Professor from UC Berkeley's Electrical Engineering and Computer Science Departments specializing in Statistical Methods and Optimization
- Eric Kansa, Adjunct Professor from UC Berkeley's School of Information specializing in open standards, public data, and the complexity of making sense of shared "raw data"
- Saheli Datta, computer assisted journalist and graduate of Columbia School of Journalism

We came to initial meetings with our ideas on tracking influence and our colleagues were able to apply expert knowledge to the discussion based on Laurent's previous efforts with the StatNews ("StatNews Project", n.d.) project and Rob's efforts with Dispute Finder ("Dispute Finder", n.d.). While neither project was a direct fit or offered reusable code, their experience was pivotal in narrowing down our line of inquiry and revising certain expectations to fit our timeline.

Ongoing efforts with these stakeholders focused on idea formulation, concept iteration, and available statistical/computing methods we could use to accomplish our tasks.

One early suggestion was to control the volume of our index and it was suggested that we focus on the health care debate of 2009. This area gave us controlled scope that was highly discussed over a concentrated period of time. Initial ideas involved doing organizational/individual wiki-like pages that highlighted shared features along lines of parallel language. The idea was to illustrate the correlations among organizations via commonalities across categories.

Another suggestion was to pick a single topic and build a system that aggregated and linguistically processed all content on this one topic space for a "daily view" of all relevant linguistic patterns on the subject.

For technical matters we began testing methods and ranking, and met with the team regularly to iterate over our process and show results. The input of this group was essential in developing the final system described in Sections 6.5 and 6.6.

5.2. Iterative Problem Definition and Troubleshooting

Based on our interviews with stakeholders and ongoing research we modified our approach iteratively, recognizing new pathways and objectives.

Our discussions with investigative journalists helped to educate us more richly on the tacit knowledge factors that veteran journalists apply to their work. At first we were somewhat discouraged as we felt it would be difficult to find a way to interest traditional journalists in these tools, but over time we began to see how a properly scoped tool which heavily stressed the needs of human intervention, could be used to augment human processes. These discussions had direct application on the interface design choices applied throughout the project and especially with regard to Section 6.5.

Finding expected patterns manually to validate potential

To test our early ideas we performed manual searches with controlled language samples. Trying to mimic journalistic intuition, we scanned PR feeds and found what we felt was constructed language that was appealing to a line of inquiry. From there we migrated our search to the aggregated LouisDB search to make a correlation across data sets. The following three results met criteria of being interesting, having a back-story, and implying a potential timeline based narrative:

i. "Health Care Takeover"
ii. "H1N1 Vaccine"
iii. "Conscience Protection"

At a later stage we applied the same type of test using our system to locate language cited by Robert Pear in his story. Specifically, comments attributed to Representatives Joe Wilson and Blaine Luetkemeyer referring to "biotechnology (as)... a homegrown success story that has been an engine of job creation in this country"(cite). Pear's requirements probably didn't require that he list more than two reuses of the language, but we were able to find seven instances in close sequence to each other, which for us, as outsiders was quite informative of the process of language reuse.

Manhole Barrier Security Systems Use Case

Another parallel scenario was stimulated by the discovery of a novel position paper from the a security organization/company 'Manhole Barrier Security Systems' (MBSS). Search-

ing for their work via Google revealed a PRNewswire document they submitted in 2008, calling for increased security for underground infrastructure and making correlations between prior terrorist activities and current vulnerabilities (cite:PrNewswire, 2008). Donations tracked by opensecrets.org illustrate that MBSS paid lobbyists approximately $100,000 in funds per year between 2006 and 2009 ("Lobbying Spending Database", 2010). In 2009, MBSS was in the news as being linked to allegations that "U.S. Rep. Pete King funneled $3 million in taxpayer money to a campaign donor for custom manhole covers" (cite:Lesser, 2009).

Once our tool was finalized, for illustration we ran a block of MBSS text through our system and found the following top fifteen language correlations.

1. Point to underground
2. Underground in nearly
3. Public utilities and telecommunications
4. Easily disrupt
5. Housed underground
6. Damage with considerable
7. Common avenue
8. Lies underground
9. Accessible through one
10. Conduit that transport
11. Provides terrorists
12. Terrorist attacks around the globe
13. Citizens take for granted
14. High among these priorities
15. Protect civilian populations

5.3. Prior Art that Effected our Research and Discussion

Text retrieval and processing are enormously popular domains and much has been done to automate the identification of certain types of language, and to apply NLP constraints to service a particular domain. We will briefly review a few instances of Prior Art that have been influential on our thinking as we have approached this problem.

Current Tools for Interacting with Legislative Documents

As legislative documents and in particular, the Congressional Record are a cornerstone of our research, we quickly focused on tools that took advantage of features of this data stream. Even more important were tools which remixed language output to give users new ways of interacting with the corpora.

Thomas.loc.gov (http://thomas.loc.gov/):

The Library of Congress' search engine for citizen interaction with the legislative documents. It offers basic browsing and advanced search functionality so that users discover

content within a specific category or via a direct query. While offering the most current data available, the system feels stale and presupposes explicit user input to drive inquiry. Each query offers a list of results, but they cannot be viewed in parallel and it lacks any power to visualize or correlate the discovered data into a context of high value for the user.

Capitolwords.org (http://capitolwords.org/):
Provides basic word frequency tabulation of new text as it enters the Congressional Record and offers users a quick entry point to view the language during a particular time period. Adding context like topic, speaker, region, the user can get a basic reading of what the Congress is up to and see what political agendas are most dominant. While a very simple tool, it offers categorization and visualizations that provide an interesting layer of value. Our assessment of the service was that it while novel, it did not offer tools for deeper correlations across multiple corpora, and the limitation of a single word, while able to show the most general of trends, lacked the ability to offer greater context or even direct the enquiry toward the source document.

Govpulse.org (http://govpulse.us/):
Does not necessarily focus on language, but rather offers metadata based organization of legislative documents via an enhanced UI that improves user interaction with legislative data. It allows easier discovery and improved lateral navigation across parallel themes. The system offers some topical and regional visualizations that can help the user get a high level view of what organizations and government projects are being most regularly cited in the CR. Overall, Govpulse offers a logical extension of legislative documents and is probably more in line with what Thomas.loc.gov could offer to improve user experience, however it doesn't offer any significant power level tools for searchers to get significantly increased value from what was already in the Thomas.lov.gov.

Memetracker
A recent news and NLP project undertaken by Cornell and Stanford, Memetracker (Leskovec, 2008) offers visualization of key language memes and their variations over time. Scanning a huge sources of news media, this NSF and Google sponsored project was able to identify the transfer of pre-defined memes from mainstream text to the blogosphere and to observe the general lag between proactive news media and reactive news blogs. The primary aim was finding unique and "contiguous sub-sequence(s) of...words in (each) phrase"(3), clustering them, weighting terms, and then using weighting to edge link these phrases to each other. The system provided a dynamic Flare visualization of the "temporal dynamics"(5) of these phrases during a fixed window. The project is no longer aggregating stories or performing any real-time analysis.

A fundamental difference between the Memetracker project and our emphasis is Memetracker's dependence on quoted text as the starting point for further analysis. Within Memetracker's focused sources of news media, quotes are a common attribute and easy to extract. By contrast, our sources are often devoid of explicit quotes, but rather is voiced at the organizational level or at the author/source level. With the government documents we can sometimes identify the speaker or commenter, but ultimately our interest is in the broader context of the affiliated quote and its connections. Our estimation was that we would find some memes, but this would not be our only goal as we are more interested in

broader knowledge transfer and though memes are interesting, they have limited a function by themselves.

Plagiarism Detection

Prior efforts in the domain of plagiarism detection also offered a narrowing focus for us to apply to our thinking. In particular, the overall degree of similarity between multiple documents helps to expose direct copying that occurs, but might also suggest authors or organizations under the influence of other authors or organizations.

Given the proliferation of digital documents it is easier than ever to duplicate text indiscriminately, whether intentional or not. Detection systems have been developed that apply different language processing techniques to a set of documents in an attempt to determine if language from one source was used in another.

Technical Foundations of Plagiarism Detection

There are many plagiarism detection techniques, but they fall mainly under two categories: techniques for comparison of a single document in a corpus to all of the others in the total corpus, or techniques for comparison of one small set of documents to one another (Lukashenko et al, 2007, p. 3). Commonly used algorithms include calculating the cosine similarity between document vectors, as well as techniques involving matching of similar sequences of sequences of words, often referred to as "n-grams" (p. 3).

Some researchers have asserted that "plagiarism is particularly difficult to test for," using only a computational system, and that the detection of identical text between two documents requires the judgment of a human being before an author could be accused of plagiarism (Brin, Davis, Garcia-Molina, 1995, p. 15). Ryu, et al discuss the potential for difficulty in determining the original date that online documents are published, and by extension, the difficulty in determining which source appeared first (Ryu et al, 2009, p. 5). Our project addresses this issue by attempting to index documents that ideally have a clear date of publication however not every document in our index has this feature.

The Meter Project

Developed as a collaboration between the University of Sheffield and the British Press Association (PA). The function of this research was to "investigate how text is reused in the production of newspaper articles from newswire sources and to determine whether algorithms could be discovered to detect and quantify such reuse automatically (Gaizauskas, 2001, p. 1). The study had structured data provided by the PA and unstructured data from nine British newspapers. The index was further restricted to court and entertainment news only, and spanned a total of 1716 stories.

Ultimately, the Meter project relied on "n-gram overlap measures, Greedy String Tiling, and sentence alignment" to compute three vectors of similarity. Of the 1716 stories, the system was able detect a 77% rate of reuse at some level of similarity. (Clough, 2002, p. 4) Among the results, the most interesting to our project was the accuracy of overlapping rare 1-grams (or single words) as the best predictor of text re-use. From this baseline, we were able to recognize that clustering candidate document pairs along this metric would yield interesting results.

6. Technical Approach

6.1. Data Acquisition

Our goal was to acquire, parse, and convert source documents into a structured, standardized, and machine-readable format suitable for indexing and text retrieval.

Our first challenge was to acquire our content from a diverse set sources and to merge these sets to a cross-functional data model. Under ideal conditions, data formats feature structural elements that enrich the document text with additional, machine processable functionality. However, online data is often published in formats meant strictly for presentation, and not well provisioned for computational parsing. For example, the structural cues that are used for visual presentation (such as font formatting markup) can be problematic for parsers that aim to extract blocks of raw text for language analysis.

Content Extraction

Acquiring diverse content from multiple sources

We created a standard data model, or schema, to describe the source documents in our index. This data model included information about the publishing, the document's source URL and title, and when possible, information about the author and date of publication. There were some corpus-specific fields such as geographical region for PRNewsWire corpus and congressional metadata from the legislative documents corpus. While our index contains considerable metadata, we were not able to collect consistent information about every document. (See Appendix 2)

Much of the source material that we indexed originated as online files in HTML format. HTML is a presentation format that contains both text as well as markup that describes the layout of the text. While there are opportunities for this format to contain information about the document's content (such as author, published date, etc.), HTML does not enforce a strict standard for this type of data. For example, when parsing HTML, the element name containing the main content differed widely across different corpora. Our team had to apply sizable efforts to figure out which tags should be parsed and saved. For example, one corpus placed their main content within while another used . Acquiring metadata was another challenge. Where one source provided a HTML META tag for publication date, another might embed the date of publication in part of the HTML URL. In total, it was necessary to write custom parsing scripts for each document source publisher.

Other documents in our index were available online in Adobe's PDF format. As a fixed-layout presentation format, PDF is often used to guarantee consistent presentation on all displays. This encoding required translation using a custom script based on the open-source Python library "PDFMiner." ("PDFMiner", n.d.)

While PDF documents also contain mechanisms for storing metadata such, there was no guarantee that this metadata was embedded with every document. Undeterred we were often able to extract additional metadata on the accompanying web pages which linked to the PDFs. The results of this process were mixed, but we were often able to extract publication date and author information.

Challenges

As described, acquiring data necessary for this project was a major challenge, which resulted in considerable investments of labor and time. The main issues contributing to this difficulty were:

i. Lack of machine readable formats for some corpora
ii. High format variability
iii. Inconsistent metadata practices and a lack of describing information in the document itself

The PDF format also contains decorative elements, multi-column layouts, and other format styles that can insert code noise that was unconvertible via PDFMiner. Content structuring creates noise and available programmatic conversion tools do not always produce reliable or expected output. One of our data sources, the National Bureau of Economic Research (NBER), encoded a large batch of PDFs in non-standard format that our PDF parsing library could not reliably turn into raw text, resulting in approximately 20% data loss. This was the only corpora that suffered such a significant extraction problem.

The resultant text output from our conversion step did not always produce results that could be completely indexed, and included non-ascii symbols, lists of single character "noise," and groups of words from block sections of text that were arranged in columns. Our PDF processing scripts would often not properly recognize the spacing between letters in a words, returning two word fragments broken apart in the middle. It was not uncommon to observe words such as "government" to be parsed as "govern" and "ment." These resulting nonsense words caused two problems. First, they affected the total word counts of our index. By reducing the number of times a word was recognized by our computational analysis, made the word slightly more rare than it was in actuality. More importantly, the nonsense terms could prevent some potentially interesting phrases from being recognized, if the broken words create fragments inside a specific phrase in a particular document.

Structuring Data

Storing

During processing, it was necessary to temporarily cache each source file on our server for indexing. The content from these documents alone required tens of gigabytes of storage. To store our index of document metadata we used the popular open source database solution, MySQL for data portability and ease of use. Our data model was designed to be flexible and allow for additional data descriptors to be added in the future.

Linking

For our system to act as a tool that attempts to indicate patterns of language influence, it was necessary to be able to reference the original source document, as well as to provide

data about the occurrences of a piece of language over time. For each document, we stored a unique identification number, a unique filename, the source of the document, and most importantly, a URL where the publication could be retrieved online. We also stored a cached copy of the raw text content of every document for which we had the legal right to store copies.

Data Modeling

While our entire document index was created using scripts written using the scripting language Python, most of the sources we accessed required custom programming to extract data from the disparate formats used by our data sources. Ideally, source documents should be defined by a standard data format that enforces a minimal but mandatory amount of metadata.

Another consideration for future data acquisition would be to design our system in a way that allows outside users to input their own corpora. While our current data model is both simple and complete enough to accept documents with minimal metadata, the main technical challenge would focus on either being very flexible in the type of data formats allowed at input, or to aid users explicitly in self-structuring their data to properly interact with the system in a reliable way.

Universal Challenges

Bandwidth Limitations

Providing the bandwidth for transmission of many gigabytes of content over a network can be costly and time consuming for both the content provider and the researchers collecting the data. Even if the time needed to access the content of a single document stored on the web is only a few seconds, the total time of accessing hundreds of thousands of documents can be a major logistical challenge. The solutions to network issues may include designing systems that are able to access and index online documents overnight without supervision, or by compressing files before transfer over a network.

Noise

Due to the limitations of automated parsing and indexing as discussed above, our document index contained some word fragments and noise. This impacted our project by affecting the statistical methods used to score the relevancy of extracted terms, which depend on the frequency of each individual word, as well as the total word count of the corpus (see 6.4).

Final Assessment of Indexed Corpora

Another consideration to note is the possibility that the size of our document index was not ideal to pursue our research goals. One of our initial assumptions was that the detection of influential language was a “needle in a haystack” problem, in which there would be a benefit in collecting the maximum amount of data before analysis. However, the investment of effort in large scale data acquisition ultimately reduced the amount of time and processing resources available for analysis and refinement of our methods.

However offering an appropriately sized dataset was an important expectation in being able to address broader research questions for computational journalism. To illustrate this point, consider two computational journalism projects that feature very different corpus

sizes. The Meter project, as discussed above, focused on developing algorithms to automatically track how language is reused in newspaper articles. The Meter corpus was relatively small (under 2,000 articles) in part because the project required that each document in the corpus be "classified by an expert journalist" (Gaizauskas, 2001, p. 6). The Memetracker project, in contrast, follows the mutation of quotes and phrases in the news media over time (Leskovec, 2008). This project makes a point of visualizing and comparing the most popular common quotes and phrases across as many sources as possible. To determine which phrases are the most popular, Memetracker claims to analyze the text of "900,000 news stories and blog posts per day from 1 million online sources."

Unlike the two systems discussed above, our focus was not to study the most popular phrases in our data set, nor was it to simply detect language reuse. Rather, we were chiefly interested in detecting expressions of influence through observing language transfer. Therefore, it was important that our data set reflect a wide array of content sources. To be able to have a large enough data set to detect the rare terms repeated across several corpora, we attempted to index as many documents as possible from each data set. However, it may have been possible to fully address our research question using a smaller data set, by more strictly limiting ourselves to a specific policy issue. This was initially suggested but our ambitions exceed what was perhaps most reasonable given our time constraints.

6.2. Computational Considerations

As mentioned our project required significant computational resources to process the quantity of source material we collected for our analysis. In particular, the following tasks were especially computationally intensive:

i. Caching and indexing large amounts content from various online sources
ii. Analyzing cached files for statistically relevant phrases
iii. Storing an index of document metadata and extracted phrases
iv. Providing users with online access to these resources

Methods

Storing in Memory

With an index of well over one million files, our computational analysis techniques pushed the limit of our available computational resources. Our current server featured the maximum allowable memory of 6 gigabytes. However, the total size of our cache file proved to be much greater that this limit. Therefore, it was impossible for us to store the complete index of file content in our server's memory at any one time. These limits forced us to carefully design our scripts to process data in batches small enough to fit within our available memory.

Bloom Filter

One technique for storing a great deal of information in a limited memory space is through the a data structure called a Bloom filter. A Bloom filter provides a mechanism to add a

large amount of elements to a set that can be later checked against for set membership. The trade-off with a Bloom filter is the increasing possibility of false positives as more data is added. While we initially used a Bloom filter in early tests, we ultimately discontinued this approach, as it proved to be unnecessarily cumbersome with respect to our goals.

Stemming

Another method of reducing data size and processing requirement in text analysis projects is the strategy of stemming. Stemming reduces words to their root form relying on common suffixes. However, since we were interested in identifying exact language matches between already rare phrases in different corpora, we felt that the benefits of stemming would be offset by the potential for false matching between stemmed words with different meanings. For example, the commonly used Porter stemming algorithm shortens both the terms "global warming" and "global warmness" to "global warm." If we used the stemmed version of either term in our index of potentially interesting n-grams, we might return false results that do not reflect phrases with similar meanings.

Available resources (Big Computers, Open Source Tools)

While the server used provided sufficient capacity for indexing the corpus text, this machine was not meant to be used as a public-facing webserver. Therefore, we decided to provide access to our index of document metadata and our index of relevant n-grams the cloud based Ruby on Rails service known as Heroku. In this model, data processed on our internal server would be pushed to our public webserver for visualization. This method of analysis was acceptable for demonstration purposes, but it is not a suitable design as an ongoing platform which requires continuous acquisition and indexing of newly published documents.

There other software projects that attempt to address these kinds of data-intensive problems via distributed computing solutions. Google's MapReduce and Apache's Hadoop are two examples that could have allowed us to spread computational tasks over a distributed network of computers. Another option for increased computational power was the Open-Cirrus cluster, a research computing cloud that we access to through our partnership with Intel Research. While these solutions provided an increased amount of processing power, each of technologies came with trade-offs that would have required significant investments of time, while reducing flexibility and not directly solving our research problem. Therefore, for our initial system prototype, we decided to simply undertake all of our processing using a single multi-core server.

Evolving the System for Future Potential

As a research tool, this system would lack utility for journalists if data was not updated frequently. Therefore ongoing viability requires the ability to continuously integrate new data into the index. Future considerations must allow for:

i. The ability of the system to handle constant database growth.
ii. Error-free automation of the system's processing and data acquisition.
iii. Reliable and fast network connection.
iv. The ability to provide enough processing capacity to index millions of documents.

While the processing capacity of a single server was sufficient for our initial proof of concept system, there are other more viable options for the project to continue in a sustainable, computationally scaleable manner. Most likely would be the use of a hosted cloud computing environments, such as Amazon's Elastic Compute Cloud, that can provide us with access to large amounts of processing and storage at relatively low cost. If these costs over time can be addressed, it might be possible to utilize a cloud computing platform to host both index processing and online database storage.

Another intriguing possibility for future computing solutions include providing an interface to an existing web scale n-gram indexing project like Microsoft Research's Web N-gram Service("Web N-Gram", n.d.). In this model, we would not be responsible for the acquisition or indexing of online content, but rather the interface for users to study potentially influential language within a index of documents. This approach has the drawback of not allowing us to rigidly control the size of the indexed n-grams, but it may provide a more viable solution to the problem of constant data acquisition and growing computational needs.

6.3. Language Preprocessing

Computational Normalization of Cached Text

As our document index was very large, we took steps to normalize the terms in our index to reduce the computational complexity. These normalization steps included:

i. Converting all processed text to lower case form
ii. Removing some punctuation from sentences before parsing
iii. Disregarding phrases with excessive noise

Determining word counts of each word in the entire corpus

Once we determined how cached text would be normalized, we calculated the overall word count of each term. This was accomplished by reading the raw text contained in each cached source file, normalizing the text according to the methods described above, and then adding the count of each individual word to a count (hash) kept in memory.

Our initial analysis strategy ranked phrase rarity by using the individual term frequency of each word in the phrase (see Section 6.4). Very common words contribute negatively to this phrase ranking. Noise or nonsense words are generally random and thus repeat infrequently, meaning that we were able to filter out noisy phrases based on a minimum term frequency of each word. In general, the overall count of these non-word terms was typically very small and the likelihood of reoccurrence across multiple corpora meant that these terms would be ignored in the final results. Therefore, we did not need to use a stop word or dictionary list to verify the validity of each word, and all collections of characters, numbers, web addresses, serial numbers, and other non-word terms were counted despite their minimal value.

Creating an index of the terms in our source document

Once we processed text into a normalized format, the cached text was indexed based on the frequency of each term in each document. This step was necessary for later retrieval of information about which document contained the which language.

For the task of indexing for phrase retrieval we chose software from the Lucene project ("Apache Lucene", n.d.), an open source indexing and search tool optimized for fast retrieval. Lucene provides several benefits, such as stability, a large user base and support network, and the ability to quickly index a large number of cached documents.

Extracting and indexing phrases from cached text

Next, we extracted collections of sequential words, commonly referred to as "n-grams", from our cached raw document text, creating another index. An n-gram is simply a sequence of words in any particular order, where "n" designates the amount of words in the set. Therefore, the phrase "four score and seven" is known as a "4-gram." In a given document of ten words, there are seven distinct 4-grams.

N-grams of word lengths between 2 and 8 were extracted from the previously cached and normalized raw text. To accomplish this we broke raw text into sentences by splitting the text on periods by using the "sentence_tokenizer" method from a Python library, the Natural Language Toolkit (NLTK). Using this new list of sentences, we looped through each sentence and retrieved the collection of all n-grams between length 2 and 8, using NLTK's "word_tokenizer" method. With pre-determined criteria described in Section 6.4, we selected relevant n-grams and stored them in a new index. Once our list of relevant n-grams was complete, we ran a query per n-gram against our document index to determine which documents contained each n-gram. We stored this relationship map in a MySQL table that could be quickly accessed to provide data for visualizations of phrase co-occurrence.

Challenges in the Extraction Stage

The choice of what length of n-gram to index was influenced by a number of factors, including our processing and storage capacity, as well as our project goals. Indexing every possible n-gram of every possible length (the longest being the longest sentence available in our cached text) could be useful for some applications (such as plagiarism detection), but may be unnecessary for detecting rare terms that appear in documents from multiple sources. Also, storing such an index of n-grams would require a great deal of storage space, as well as additional processing power to generate query results. Therefore, we focused on indexing only the rarest short n-grams in our data set, in order to maximize the utility of our limited computational resources.

As previously discussed there were ongoing challenges with improperly parsed words. For example, if a think tank provides its publications in PDF format, and each document features a recurring word that is often parsed incorrectly, then it could introduce noise which was deemed relevant and rare. Consider a word that might appear in the text of many documents such as "page" (as in page 1, page 2, etc.). If this word appears in these documents using a font or style that complicates the task of extracting the text programmatically, it may appear in our index as multiple occurrences of the nonsense words such as "pag" and "e."

In contrast to the possibility of noise in our cached text collection, there may also be intentional hyphenation based on formatting considerations. For example, the word "congressional" might be written as "con-gressional" as it crosses a line break. Each fragment would be recognized by a computer parser as an independent term. This would result in a reduced term frequency of the word "congressional." While the amount of hyphenated word fragments is likely to be low compared to the amount of whole representations of the same word, it is important to note that our system can not currently address this discrepancy.

6.4. Language Pattern Discovery

> *We are, as language animals, to a substantial degree simply permuters of substrings already well established within the history of the whole language (Wilks, n.d.).*

An stated, this project seeks to construct a system with the ability to retrieve units of text that might interpret as influential. Therefore, our starting task was consider the structural components that would be common in language that was uncommon, or by extension, laboriously constructed by the writer to impart a particular agenda. Initial starting points included:

i. Sequences of words into combinations that are rarely found together;
ii. used relatively infrequently within the overall index of language;
iii. appearing in multiple documents from different sources.

Tuning the System to the Task of Finding Unique Language

In our discussion of rarity, we consider how likely terms are to occur together in a given document. We view these as statistically improbable and significant. Our goal is to distinguish these from noisier combinations such as acronyms which occur frequently or proper names which may occur infrequently.

Subsequently, depending on the requirements of the user, there can be different value in shifting the baseline calculations of rarity. For example, a phrase from Shakespeare would appear very out of place in the Congressional Record. Similarly, phrases like "Madam I rise" are common in the Congressional Record but very unusual in text written by think tanks. By choosing the appropriate baseline language set, we are able to improve the results depending on the corpora of interest to the user.

Ascertaining Language "Rarity"

To determine how unlikely a particular phrase is, a score consisting of the count of each word in the phrase is divided by the total word count of the index. Next, the log of each scored word is summed with other words in the n-gram. The lower this summed score is, the rarer the phrase is in relation other n-grams of the same length. This approach is an approximation, and assumes no particular rules of grammar besides word frequency.

Document Co-Occurrence

Once an n-gram is identified as potentially significant according to the rarity test described above, it is then queried against our index of documents. This query returns references to the list of documents in which the n-gram can be found. Besides providing a mechanism for linking n-grams back to their source documents, this count can be used as yet another method of scoring n-grams. For example, a statistically rare n-gram that appears in two documents from different sources might be a candidate for further review as a unit of potentially influential language. Furthermore, once the relationship between an n-gram and a document is determined, it is possible to provide statistics about which n-grams tend to co-occur together within documents.

Qualifying Influential Transfer

Another feature we are looking for is influential language that crosses corpora boundaries. For example, a rare phrase that appears in a position paper by a policy think tank in September of 2006, and then appears again in the Congressional Record in December of 2006, might be a potential candidate for influential language.

Limitations and Next Steps

It is important to note that, while we feel that our strategy produced a suitable proof of concept, it is far from a complete system for finding influential language. We are suggesting a method to extract certain unique language, and we are applying a method to watch the transfer occur, but there are human language filters which can be applied that are far superior to computational ones available to us. For example, while a term may be rare, it is not necessarily going to be interesting to humans, nor explicitly indicative of influential language. What we can do is decrease the visibility other non-rare language, or non-multi-corpora language, helping the user get a more focused view. With new methods for natural language processing, computational systems may become more accurate, but for now we see this as a collaborative system which augments human finding abilities.

6.5. Graphic Interfaces for Interacting with Data

User interfaces were designed chiefly to empower users to be able to drive an inquiry against the datasets. The thinking was providing multiple points of entry to the system could allow users to either “investigate” or “explore”. With the “investigation” interface users bring self-identified text to the system via cut/paste interaction. Unlike in the Robert Pear case, they do not need to intuitively focus on small blocks of interesting language. Instead the system allows for a very large blocks and it identifies the possibly constructed language automatically. The system provides these in a ranked order, and the user can augment the process by narrow down the result set using metadata, or by diving further into a single n-gram to look at visualizations or context. With the “explore” interface, users would enter the system from general positions (i.e looking at top level corpora, predefined lists of Top 10 n-grams, or user favorites).

Other important visualization modalities:

Timeline: How term patterns emerge over time

Show the relationship of language across corpora in an intuitive temporal interface. Visible as a view for any n-gram to illustrate the context of the n-gram over time. Multiple n-grams can be viewed in timeline mode while stacked on top of each other in n-gram list mode.

Each document will be visible as a node on the timeline and color coded to reflect the corpus it is from. Nodes will be clickable and will illustrate textual context view. Object properties will be expressed via a metadata view that show document title, author, organization and date.

Context view

To add depth of reference to multiple occurrences of the same n-gram. Will allow user to quickly jump to read the 5 sentences that exist in relation to the selected n-gram. Similarly, the user can quickly select new document instances of the current n-gram to read in parallel, or can jump to a new n-gram context view.

Top ten N-Grams for each Organization in the Corpus with N-Gram color coding

Illustrative and exploratory mode to see the top ten most unusual language located in each corpora. Allows users to get a overview of language differences between the different institutions and to dive in to n-gram view with context when desired.

Metadata as a Mechanism to Constrain Results

Not implemented universally given disparate metadata, but minimal options will allow users to constrain results by date range, n-gram length, organizational or corpus level attributes. Will help users to increase the strength of filtering to clearly view specific context.

6.6. Final System Implementation

The interface to our prototype system was designed with the following use cases in mind:

i. Users who are interested in finding potentially influential language in some sample text
ii. Users who want to learn about the propagation of a term over time across corpora
iii. Users who are interested in studying the statistics of our entire database

Exploration and Investigation

In Migratory Words' Investigation mode, the system accepts user supplied text, and attempts to provide the user with a set of phrases that appear in both the query text, as well as phrases that appear in five other documents in our index that are similar to the query. Behind the scenes, the system compares the query text with our document index, returning a set of phrases that appear in the query as well as in five other documents that have the closest term vector similarity (using Lucene's MoreLikeThis method).

Once a set of interesting phrases are returned, the user is able to enter Migratory Words' Exploration mode. The user may interact with the phrases of their choice by clicking on

the phrase text to reveal a list of documents containing that and visualize the relationships between phrases and documents. Migratory Words also allows the user to view a snippet of the phrase in context of any matching source documents. The user can either click to be taken to a view that allows them to dig deeper into how the phrase is used, or they can follow a link off-site to the source document's web page.

Another feature that can lead a user into our system's Exploratory mode is our list of the top interesting phrases that are in our data index. This list can be compared across corpora by referencing our database of data sources. The phrases provide another input to our system, allowing users to get a sense about what type of phrases tend to be used by which data sources.

Tracking the User's Journey

The system will be able to remember the history of phrases that they have interacted with. The utility of this feature can be expanded into visualizations about the how the phrases might be related to each other, including information about common corpora.

User Driven Scoring

We have mentioned previously that this system may be used to augment a user's tacit knowledge of a problem space. However, the system may also collect feedback about what a user might find interesting. Therefore, each phrase has a mechanism (currently a clickable icon) that allow the user to report if a phrase seems to be interesting or not. This will create will be added to a column in the database which can be used to create a User Favorite's section or to be incorporated into the scoring system to add weight to the term.

7. Results

7.1. Specific Deliverables

The list of deliverables for this project include:

Webservice Platform @ migratorywords.com
- Contains:
  - Search/Investigation interface
  - Exploration Interface
  - n-gram List View
  - Context features

Open Source Software package @
http://code.google.com/p/computational-tools-for-investigative-journalism/

This Paper
Poster
Presentation

7.2. Testing Some Results

If an uncommon term appears relatively frequently in documents from different sources, what does this mean? To test this out, we applied a search process to assess our results.

Case Study: Journalist doing research on Funding for Abortion

Scenario Basics:
User sees a PR feed posted by the US Conference of Catholic Bishops. User pastes text into the system to get a sense of the language used and to find out where political affinities lie in relation to the language used. The successful result would show the unique language used, rank in such a way that the more unusual text is singled out and the user can experience the context of usage across different publishers, and then allow the user see the language usage over time as it defines or connects with current events.

Expectations that Should be Met to Qualify Success:
- The rarity of language should reveal highly constructed, message based language.

- The ranking system should provides sensible sorting to help the most unique language bubble up.
- The context provides a way for the user to get greater depth on the language as it is provided in relation to the same usage in other published material.
- The source of the document publication can show coordinated or uncoordinated views between organizations. Also, it with context the user can see how organizations support or dispute the same idea.
- The timeline may show significant movement in concentrated time. The first instance may suggest the origination point of the language usage (in the corpora). The change in frequency could suggest growth in influence, or irregularly heavy usage during certain periods of great debate.
- Document may show transfer from advocacy group to government text.

The initial input was a block of text from a United States Conference of Catholic Bishops press release entitled "Bishops Call Senate Health Care Reform Bill 'Deficient,' Essential Changes Needed Before Moving Forward" from Dec 22,2009.

The block of text used to query Migratory Words was:

> *They said that the health care bill passed by the House of Representatives "keeps in place the longstanding and widely supported federal policy against government funding of elective abortions and plans that include elective abortions" and "ensures that where federal funds are involved, people are not required to pay for other people's abortions." The Senate bill does not maintain this commitment.*
>
> *In the Senate version, "federal funds will help subsidize, and in some cases a federal agency will facilitate and promote, health plans that cover elective abortions," the bishops said. "All purchasers of such plans will be required to pay for other people's abortions in a very direct and explicit way, through a separate premium payment designed solely to pay for abortion. There is no provision for individuals to opt out of this abortion payment in federally subsidized plans, so people will be required by law to pay for other people's abortions." The public consensus against abortion funding, said the bishops, "is borne out by many opinion surveys, including the new Quinnipiac University survey December 22 showing 72 percent opposed to public funding of abortion in health care reform legislation."*
>
> *"This bill also continues to fall short of the House-passed bill in preventing governmental discrimination against health care providers that decline involvement in abortion," the bishops said. And it also "includes no conscience protection allowing Catholic and other institutions to provide and purchase health coverage consistent with their moral and religious convictions on other procedures."*

The output revealed a list of 46 phrases deemed interesting due to the statistical unlikeliness, as scored by the methods described in Sections 6.4 and 6.5. The results found 21 phrases that both appeared in our input query, and appeared in documents in our index more than once.

These phrases included:

Language	Document Occurrences	Interesting Facet
Conscience protection	53	Mentioned 29 times in 2009 (More Below)
Governmental discrimination	11	
Government funding of elective abortions	5	
Longstanding and widely supported	9	
Include elective abortions	6	5 PR Documents 1 Congressional Record Document

SEE APPENDIX 3 FOR EXAMPLES OF TIMELINE VIEW

Using the tool, it was determined that the phrase "conscience protection" was interesting, because while rare, our results page existed in a number of different document sources including Press Releases, Bills, Congressional Hearings, and the Federal Registry. The timeline view showed that the phrase could be found in our index 15 times between 2001 and 2003, after a gap of almost no occurrences between 2004 to 2008, the term appeared in our corpus 29 times in 2009.

Our system returns results indicating that, in 2009, the phrase "Conscience protection" also appeared in 7 Bills in our index. One of these includes the text of the "Patient Protection and Affordable Care Act."

Other press releases containing this phrase in December of 2009 include "U.S. Bishops Urge Senators to Support Nelson-Hatch-Casey Amendment on Health Care Reform", which features the language:

> "*We urge the Senate to support the Nelson-Hatch-Casey amendment. As other amendments are offered to the bill that address our priorities on conscience protection, affordability and fair treatment of immigrants, we will continue to communicate our positions on these issues to the Senate.*" ("U.S. Bishops Urge Senators", 2009)

Migratory Words also reports that a press release ('Americans United for Life', 2009) containing the phrase "Conscience protection" which appears in a press release from another organization, the 501c3 advocacy group Americans United For Life, which identifies itself as a "public interest law and policy organization." The language used in this press release includes this sentence:

> *First, the amendment provides inadequate conscience protection, because it does not prohibit any government entity or program (federal, state, or local) from discriminating against health care providers that do not want to participate in abortions.*

Results:

1) The rarity of language should reveal highly constructed, message based language.

Success - While the input text was rather loaded, the result set is a good return of the most similar language. Certain results like 'conscience protection' and 'religious convictions' single out the more meaningful phrases.

2) The ranking system provides sensible sorting to help the most unique language bubble up.

OK - Current document count give a reasonable result. Improvements are being made to make scoring more obvious and filtering more meaningful. Sorting should be more user focused with more metadata criterion for filtering and the ability to arrange results by doc occurrence or rarity score.

3) The context provides a way for the user to get greater depth on the language as it is provided in relation to the same usage in other published material.

OK - Still evolving in the UI, the context view definitely allows insight to see how the language is used in the documents in which it appears.

4) The source of the document publication can show coordinated or uncoordinated views between organizations. Also, it with context the user can see how organizations support or dispute the same idea.

Success - The discovery of linkage between the United States Conference of Catholic Bishops and the American's United for Life.

5) The timeline may show significant movement in concentrated time. The first instance may suggest the origination point of the language usage (in the corpora). The change in frequency could suggest growth in influence, or irregularly heavy usage during certain periods of great debate (See Appendix x).

Success - Definitely more activity during 2009 when the health care debate was in full swing and when there was disputation surrounding federal funding for abortion as part of the package.

6) Document may show transfer from an advocacy group to government text.

Mixed - Initial text in our index was actually Government records and a majority of early instances of this language are also Government records. However, 'religious conviction' is sees initial publication in 1985 by 5 separate think tanks prior to arrival in 2000 in our government documents. The problem with this sample is that it is invalidated by the lack of government documents going further back. This is one area where further testing and additional data is required.

Overall, these results are fairly good. With concentrated development, our accuracy and usability can be improved, but these are encouraging results given constraints.

Extending the Search as Might be Done by a Journalist:

OpenSecrets.org provides online access data about "the influence of money on U.S. politics." However, while OpenSecrets contains information about the financial lobbying activities of the US Conference of Catholic Bishops until 2008 ("Lobbying Spending Database",

n.d.), it does not contain any mention about the financial activities of Americans United For Life, or its so-called "legislative arm," AUL Action. While OpenSecrets currently lacks data on how the US Conference of Catholic Bishops spent money on lobbying efforts in 2009, it would be interesting to compare what statements or legislation were related to the recipients of funding from this organization. However, we assert that unlike the data that OpenSecrets provides, the patterns of language is informative in such a way as to shine light on constructed intention, rather than inferred intent from financial impetus.

7.3. Final Review and Shortcomings

Overall, we are pleased with our results. While the system is very much a work in progress, we believe we have the foundations for a novel approach to investigative research around the linguistic patterns of political and intellectual influence. Our extensible computational platform and sizable corpus provide a worthy foundation for further analysis. While we've only just scratched the surface of deeper influence patterns, this proof of concept illustrates the broader research potential of other custom tools centered on computational journalism. Specifically illustrating disputed policy networks, performing sentiment analysis, or clustering topical and organizational similarities, all founded upon these initial linguistic tools. However, our corpus must be dramatically expanded to support making any robust statements about its ability to transform current research practices. There are significant holes in the corpus that must be plugged, and we will require more complete data retrieval practices to reach this objective. Additional efforts will be forthcoming to support extending our services in whatever ways we can to increase the context surrounding the user query.

Furthermore, we must improve system performance and scale the system to be ready for wider distribution. As it stands, our server is simply not sufficient and we will need a distributed approach to increase Lucene response time. In deference to our computational challenges, we regret not having time for extensive user testing, but we feel a more complete user experience survey would be highly beneficial and the platform could benefit from concentrated UX development.

This has been a very rewarding project and everyone involved was challenged by the scope of technical requirements and the breadth of applied interdisciplinary study. Challenges such as these are very stimulating and developing this further would certainly be of significant value to all team members and hopefully a large community of users.

8. Supplemental Materials

Works Cited:

1. 2008_Global_Go_To_Think_Tanks.pdf. (n.d.). . Retrieved from http://www.sas.upenn.edu/irp/documents/2008_Global_Go_To_Think_Tanks.pdf
2. A Century of Lawmaking for a New Nation: U.S. Congressional Documents and Debates, 1774-1873. (n.d.). . Retrieved May 3, 2010, from http://memory.loc.gov/ammem/amlaw/lawhome.html
3. Americans United for Life Action: 'Senate Health Care Bill Absolutely Fails To Meet Abortion and Life... – WASHINGTON, Dec. 19 /PRNewswire-USNewswire/ –. PR Newswire: press release distribution, targeting, monitoring and marketing. Retrieved May 2, 2010, from http://www.prnewswire.com/news-releases/americans-united-for-life-action-senate-health-care-bill-absolutely-fails-to-meet-abortion-and-life-protections-79727602.html
4. "Apache Lucene - Overview ." Welcome to Lucene!. N.p., n.d. Web. 10 May 2010. <http://lucene.apache.org/java/docs/index.html>.
5. Apps for Democracy - An Innovation Contest by iStrategyLabs for the DC Government and Beyond. (n.d.). . Retrieved May 9, 2010, from http://www.appsfordemocracy.org/
6. Banfield, E. C. (1961). Political influence. Glencoe [Ill.]: Free Press.
7. Obama, B. (2009). Transparency and Open Government | The White House. The White House. Retrieved April 1, 2010, from http://www.whitehouse.gov/the_press_office/transparencyandopengovernment
8. Bikhchandani, S., Hirshleifer, D., & Welch, I. (2005). Information Cascades and Observational Learning. The New Palgrave Dictionary of Economics, Palgrave Macmillan.
9. Brin, S., Davis, J., & Garcia-Molina, H. (1995). Copy detection mechanisms for digital documents. ACM SIGMOD Record, 24(2), 409.
10. Capitol Words. (n.d.). . Retrieved February 22, 2010, from http://capitolwords.org/
11. chapter-21-how-can-think-tanks-win-friends-and-influence-people-in-the-media-crowley.pdf. (n.d.). . Retrieved from http://atlasnetwork.org/wp-content/uploads/2009/01/chapter-21-how-can-think-tanks-win-friends-and-influence-people-in-the-media-crowley.pdf
12. CITIZENS UNITED v. FEDERAL ELECTION COMM'N. (n.d.). . Retrieved April 9, 2010, from http://www.law.cornell.edu/supct/html/08-205.ZO.html
13. CITIZENS UNITED v. FEDERAL ELECTION COMM'N. (n.d.). . Retrieved May 9, 2010, from http://www.law.cornell.edu/supct/html/08-205.ZX.html

14. Clough P, Gaizauskas R, Piao S, Wilks Y (2001) METER: MEasuring TExt Reuse. Department of Computer Science Research Memorandum CS-01-03, University of Sheffield
15. ConOpsFinal.pdf. (n.d.). . Retrieved from http://www.ideascale.com/userimages/sub-1/736312/ConOpsFinal.pdf
16. Data's shameful neglect : Article : Nature. (n.d.). . Retrieved April 29, 2010, from http://www.nature.com/nature/journal/v461/n7261/full/461145a.html
17. Democracy 21 – A Report Card from Reform Groups on the Obama Administration's Executive Branch Lobbying, Ethics and Transparency Reforms in 2009. (n.d.). . Retrieved April 30, 2010, from http://www.democracy21.org/index.asp?Type=B_PR&SEC=%7B91FCB139-CC82-4DDD-AE4E-3A81E6427C7F%7D&DE=%7B4821A89A-7F6A-4B56-B282-59C6778C3FEF%7D
18. Dispute Finder : Reveal the other side of the story. (n.d.). Dispute Finder : Reveal the other side of the story. Retrieved May 10, 2010, from http://disputefinder.cs.berkeley.edu/
19. Disterer, G. (2001). Individual and social barriers to knowledge transfer. In System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on (p. 7).
20. Distrust, Discontent, Anger and Partisan Rancor - Pew Research Center. (n.d.). . Retrieved May 5, 2010, from http://pewresearch.org/pubs/1569/trust-in-government-distrust-discontent-anger-partisan-rancor
21. Dorf, M. C. (2010). The Supreme Court Rejects a Limit on Corporate-Funded Campaign Speech. FindLaw's Writ | Legal Commentary. Retrieved May 10, 2010, from http://writ.news.findlaw.com/dorf/20100125.html
22. EBSCOhost: The Think Tank Index. (n.d.). . Retrieved March 5, 2010, from http://web.ebscohost.com/ehost/pdf?vid=2&hid=107&sid=9ceb3775-49eb-4eb3-8701-b503e439f6a3%40sessionmgr110
23. Friedkin, N.E., & Johnson, E.C. 1999, Social influence networks and opinion change. Advances in Group Processes, 16:1-29
24. Fung, A., Graham, M., & Weil, D. (2007). Full disclosure: The perils and promise of transparency. New York: Cambridge University Press.
25. Gaizauskas, R., Foster, J., Wilks, Y., Arundel, J., Clough, P., & Piao, S. (2001). The METER corpus: a corpus for analysing journalistic text reuse. In Proceedings of the Corpus Linguistics 2001 Conference (pp. 214–223).
26. Gladwell, M. (2000). The tipping point: How little things can make a big difference. Boston: Little, Brown.
27. Google Cache Ruled Fair Use | Electronic Frontier Foundation. (n.d.). . Retrieved May 1, 2010, from http://www.eff.org/deeplinks/2006/01/google-cache-ruled-fair-use
28. govpulse. (n.d.). . Retrieved May 9, 2010, from http://govpulse.us/
29. Growth of the number of data.gov datasets. (n.d.). . Retrieved May 3, 2010, from http://data-gov.tw.rpi.edu/demo/static/demo-92-history.php
30. Improving Access to Government through Better Use of the Web. (n.d.). . Retrieved April 30, 2010, from http://www.w3.org/TR/egov-improving/#FIFTH-ESTATE

31. influence. (2010). In Merriam-Webster Online Dictionary. Retrieved May 9, 2010, from http://www.merriam-webster.com/dictionary/influence
32. It's Time for Cities to Protect Their Underground Infrastructure: (n.d.). . Retrieved May 1, 2010, from http://www2.prnewswire.com/cgi-bin/stories.pl?ACCT=109&STORY=/www/story/05-05-2008/0004806641&EDATE=
33. Kolak, O., & Schilit, B. N. (2008). Generating links by mining quotations. In Proceedings of the nineteenth ACM conference on Hypertext and hypermedia (pp. 117–126).
34. Lesser, Benjamin. “Rep. Pete King funneled money for 5,000 manhole covers that Con Ed didn't want.” New York News, Traffic, Sports, Weather, Photos, Entertainment, and Gossip - NY Daily News. N.p., n.d. Web. 9 May 2010. <http://www.nydailynews.com/news/2009/01/25/2009-01-25_rep_pete_king_funneled_money_for_5000_ma-2.html>.
35. Lessig, Lawrence. “Against Transparency | The New Republic.” The New Republic. N.p., n.d. Web. 9 May 2010. <http://www.tnr.com/article/books-and-arts/against-transparency>.
36. Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the Dynamics of the News Cycle. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 497–506).
37. Lobbying Database | OpenSecrets. (n.d.). . Retrieved April 26, 2010, from http://www.opensecrets.org/lobby/index.php
38. Lobbying Spending Database-US Conference of Catholic Bishops, 2008 | OpenSecrets. (n.d.). Retrieved May 10, 2010, from http://www.opensecrets.org/lobby/clientsum.php?year=2008&lname=US+Conference+of+Catho
39. Lobbying Spending Database-Manhole Barrier Security Systems, 2009 | OpenSecrets. (n.d.). . Retrieved May 1, 2010, from http://www.opensecrets.org/lobby/clientsum.php?lname=Manhole+Barrier+Security+Systems&year=2009
40. "LOUISdb.org: All Categories." LOUISdb.org. N.p., n.d. Web. 12 Apr. 2010. <http://louisdb.org/all/view.php?url=http://www.louisdb.org/documents/cr/2009/no/07/cr07no09-45.html&start-year=2009&start-month=5&end-year=2010&end-month=5&q=homegrown+success+story&qt=q>
41. Lucas, L. M. (2005). The impact of trust and reputation on the transfer of best practices. Journal of Knowledge Management, 9(4), 87-101. doi:10.1108/13673270510610350
42. Lukashenko, R., Graudina, V., & Grundspenkis, J. (2007). Computer-based plagiarism detection methods and tools: an overview. In Proceedings of the 2007 international conference on Computer systems and technologies (p. 40).
43. McGann, J. G. (1995). The competition for dollars, scholars, and influence in the public policy research industry. Lanham, Md: University Press of America.
44. McGann, J. G. (2007). Think tanks and policy advice in the United States: Academics, advisors and advocates. Routledge research in American politics, 1. Abingdon [England]: Routledge.
45. MemeTracker: About. (n.d.). . Retrieved May 7, 2010, from http://www.memetracker.org/about.html
46. Miller, D. (1994). Pierre Bourdieu and Loic J.D. Wacquant, An Invitation to Reflexive Sociology. The Australian and New Zealand Journal of Sociology. 30 (1), 98.

47. Ohm, P. (n.d.). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization.
48. PDFMiner. (n.d.). . Retrieved May 4, 2010, from http://www.unixuser.org/~euske/python/pdfminer/index.html
49. Pear, Robert. “In House, Many Spoke With One Voice - Lobbyists’ - NYTimes.com.” The New York Times - Breaking News, World News & Multimedia. N.p., 14 Nov. 2009. Web. 10 May 2010. <http://www.nytimes.com/2009/11/15/us/politics/15health.html?_r=3>.
50. Public's Priorities for 2010: Economy, Jobs, Terrorism: Summary of Findings - Pew Research Center for the People & the Press. (n.d.). . Retrieved April 26, 2010, from http://people-press.org/report/584/policy-priorities-2010
51. Robinson, D., Yu, H., Zeller, W. P., & Felten, E. W. (2008). Government data and the invisible hand. Yale JL & Tech., 11, 160.
52. Ryu, C. K., Kim, H. J., & Cho, H. G. (2009a). A detecting and tracing algorithm for unauthorized internet-news plagiarism using spatio-temporal document evolution model. In Proceedings of the 2009 ACM symposium on Applied Computing (pp. 863–868).
53. Statement by the President on the DISCLOSE Act | The White House. (n.d.). . Retrieved May 5, 2010, from http://www.whitehouse.gov/the-press-office/statement-president-disclose-act
54. Sunstein, Cass R. 2007 Republic.com 2.0 / Cass R. Sunstein Princeton University Press, Princeton : http://www.loc.gov/catdir/toc/ecip0712/2007008392.html
55. The Supreme Court Rejects a Limit on Corporate-Funded Campaign Speech. (n.d.). . Retrieved April 26, 2010, from http://writ.news.findlaw.com/dorf/20100125.html
56. StatNews Project. (n.d.). Electrical Engineering & Computer Sciences | EECS at UC Berkeley. Retrieved May 10, 2010, from http://www.eecs.berkeley.edu/~elghaoui/StatNews/
57. Think tanks - SourceWatch. (n.d.). . Retrieved February 22, 2010, from http://www.sourcewatch.org/index.php?title=Think_tanks
58. THOMAS (Library of Congress). (n.d.). . Retrieved May 1, 2010, from http://www.thomas.gov/
59. Clive Thompson on the Cyborg Advantage | Magazine. (n.d.). . Retrieved May 1, 2010, from http://www.wired.com/magazine/2010/03/st_thompson_cyborgs/
60. U.S. Bishops Urge Senators to Support Nelson-Hatch-Casey Amendment on Health Care Reform; Urge Constituents to Back... – WASHINGTON, Dec. 7 /PRNewswire-USNewswire/ –. PR Newswire: press release distribution, targeting, monitoring and marketing. Retrieved May 2, 2010, from http://www.prnewswire.com/news-releases/us-bishops-urge-senators-to-support-nelson-hatch-casey-amendment-on-health-care-reform-urge-constituents-to-back-it-78699207.html
61. USCCB - (Office of Media Relations) Bishops Call Senate Health Care Reform Bill Deficient, Essential Changes Needed Before Moving Forward. (n.d.). United States Conference of Catholic Bishops. Retrieved May 2, 2010, from http://www.usccb.org/comm/archives/2009/09-267.shtml
62. United States Code: Title 17,107. Limitations on exclusive rights: Fair use | LII / Legal Information Institute. (n.d.). . Retrieved May 2, 2010, from http://www4.law.cornell.edu/uscode/17/107.html

63. U.S. Senate: Legislation & Records Home > Lobbying Disclosure > Lobby Disclosure Act - TOC > Section 2. (n.d.). . Retrieved February 26, 2010, from http://www.senate.gov/legislative/Lobbying/Lobby_Disclosure_Act/2_Findings.htm
64. Web Robots Pages. (n.d.). . Retrieved May 1, 2010, from http://www.robotstxt.org/faq/legal.html
65. Web N-Gram - Microsoft Research. (n.d.). . Retrieved May 2, 2010, from http://web-n-gram.research.microsoft.com/info/
66. Weber - Politics as a Vocation. (n.d.). . Retrieved April 1, 2010, from http://media.pfeiffer.edu/lridener/dss/Weber/polvoc.html
67. White Papers - TheHill.com. (n.d.). . Retrieved February 22, 2010, from http://thehill.com/resources/white-papers/
68. White Papers - TheHill.com. (n.d.). . Retrieved May 1, 2010, from http://thehill.com/resources/white-papers/organization/434-manhole-barrier-security-systems-inc-mbss/
69. Weidenbaum, M. L. (2009). The competition of ideas: The world of the Washington think tanks. New Brunswick, N.J.: Transaction
70. Wilk, Y. (n.d.). ON THE OWNERSHIP OF TEXT. Retrieved May 8, 2010, from http://www.dcs.shef.ac.uk/~yorick/papers/cs_00_08.html

Appendix I

Corpus	size of metadata (MB)	size of content (MB)	# of documents	size ratio content/meta	avg metadata size (kB)	avg content size (kB)		meta vs. content diff
aei	78	150	19851	1.92	3.93	7.56		v
brookings	4.8	33	1222	6.88	3.93	27.00		
cato	24	91	6236	3.79	3.85	14.59		
cfr	95	180	11164	1.89	8.51	16.12		v
cgdev	1.5	26	386	17.33	3.89	67.36		
civitas	0.172	5.9	43	34.30	4.00	137.21		
fraser	7.8	153	1915	19.62	4.07	79.90		
nber	63	1100	16109	17.46	3.91	68.28		
ncpa	3.5	31	898	8.86	3.90	34.52		
rand	14	57	3669	4.07	3.82	15.54		
urban	18	258	4556	14.33	3.95	56.63		
thehill	2	27	500	13.50	4.00	54.00		
afp	188	6400	1592309	34.04	0.12	4.02		
apw	264	10000	2272995	37.88	0.12	4.40		
cna	11	339	85600	30.82	0.13	3.96		
ltw	28	1700	295224	60.71	0.09	5.76		
nyt	185	10000	1655279	54.05	0.11	6.04		
xin	143	4800	1247039	33.57	0.11	3.85		
louisdb/bills	325	1600	83115	4.92	3.91	19.25		v 83117
louisdb/cr		3100	228134	#DIV/0!	0.00	13.59	size = meta + content in xml format	
louisdb/fr	1100	3200	258464	2.91	4.26	12.38		v 26xxxx
louisdb/hearings	26	1300	6325	50.00	4.11	205.53		v 6700
louisdb/presidential	9.2	12	2346	1.30	3.92	5.12		v 2347
louisdb/reports	23	711	5942	30.91	3.87	119.66		v 5999
prnewswire0607	1600	3000	420384	1.88	3.81	7.14		
prnewswire09	350	504	89634	1.44	3.90	5.62		
TOTAL	**4563.972**	**48777.9**	**8309339**	**10.69**	**0.55**	**5.87**		
LouisTotals	1483.2	9923	584326					
TTtotals	309.772	2084.9	66049					
PRtotal	1950	3504	510018					

Appendix 2

	A	B	C	D	E	F	G	H	I	J	K
1	corpus	doc_id	trackback_	published_time	publisher	author	title	keywords	categories	description	type
2	bills	[filename]	<id>	<published>			<title>				bills
3	cr	[filename]	<id>	<published>			<title>			<description>	cr
4	fr	[filename]	<id>	<published>			<title>				fr
5	hearings	[filename]	<id>	<published>			<title>				hearings
6	presidential	[filename]	<id>	<published>			<title>			<description>	presidential
7	reports	[filename]	<id>	<published>			<title>			<description>	reports
8											
9	aei	[filename]	URL:	DC.date:	publisher:	author:	DC.title:	keywords:	taxonomy:	DC.description:	mediaType:
10	cato	[filename]	URL:	date: or date_parsed: (first)		author:	title:	keywords:	sector:	description:	type:
11	cgdev	[filename]	URL:	date:	copyright:	authors:	title:	keywords:	sector:	description:	
12	fraser	[filename]	URL:	Date Published:		Author(s):			sector:		Type:
13	thehill	[filename]	URL:		organization:		title:	keywords:	sector:	abstract:	
14	brookings	[filename]	URL:	date:	bi_center:	bi_author:	title:	keywords:	bi_taxonomy:	description:	bi_doctype:
15	cfr	[filename]	URL:	date:	publisher:	authors:	title:	keywords:	issue: (multiple)	description:	pubtype:
16	civitas	[filename]	URL:	date:		author:	title:		sector:	description:	
17	ncpa	[filename]	URL:			authors:		keywords:	type:	description:	
18	rand	[filename]	URL:	DC.Date:	DC.Publisher:	DC.Contributor: (multiple)	title:		DC.Subject	description:	type:
19	urban	[filename]	URL:	DC.date:	DC.publisher:	Author:	DC.title:	Keywords:	DC.subject:	DC.description:	DocumentType:
20	prnewswire0607	[filename]		date:		CO:			IN:		
21	prnewswire09	[filename]	separate urls file	date:	publisher:	author:	separate urls file	keywords:	categories:	description:	
22	nber	[filename]	URL:	DC.Date:		author: (multiple)	title:	keyword: (multiple)	sector: (multiple)		type:
23											
24	afp/apw/cna/ntw/nyt/xin	id		extracted from "id"			HEADLINE				type

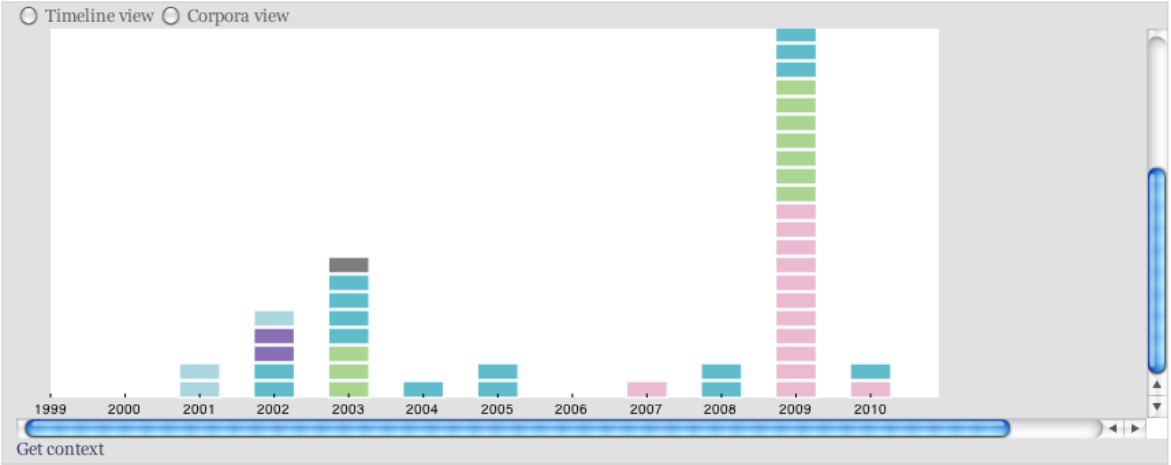
L	M	N	O	P	Q	R	S	T	U	V
geo_region	geo_place	brooking_pro	prnew	prnewswire_stock	louisdb_congress	louisdb_session	louisdb_volume	louisdb_section	louisdb_chamber	louisdb_type
					<louisdb:congress>	<louisdb:session>				<louisdb:type>
							<louisdb:volume>	<louisdb:section>		
							<louisdb:volume>	<louisdb:section>		
					<louisdb:congress>				<louisdb:chamber>	
								<louisdb:section>		<louisdb:type>
					<louisdb:congress>				<louisdb:chamber>	<louisdb:type>
		bi_program:								
region:										
ST:			SU:	STK:						
geo_region:	geo_place:									

V	W	X	Y	Z	AA	AB
louisdb_type	louisdb_number	louisdb_version	louisdb_startpage	louisdb_endpage	louisdb_speaker	dateline
<louisdb:type>	<louisdb:number>	<louisdb:version>				
			<louisdb:start-page>	<louisdb:end-page>	<louisdb:speaker name-id="(.*)">	
			<louisdb:start-page>	<louisdb:end-page>		
<louisdb:type>			<louisdb:start-page>	<louisdb:end-page>		
<louisdb:type>	<louisdb:number>					
						DATELINE

Appendix 3

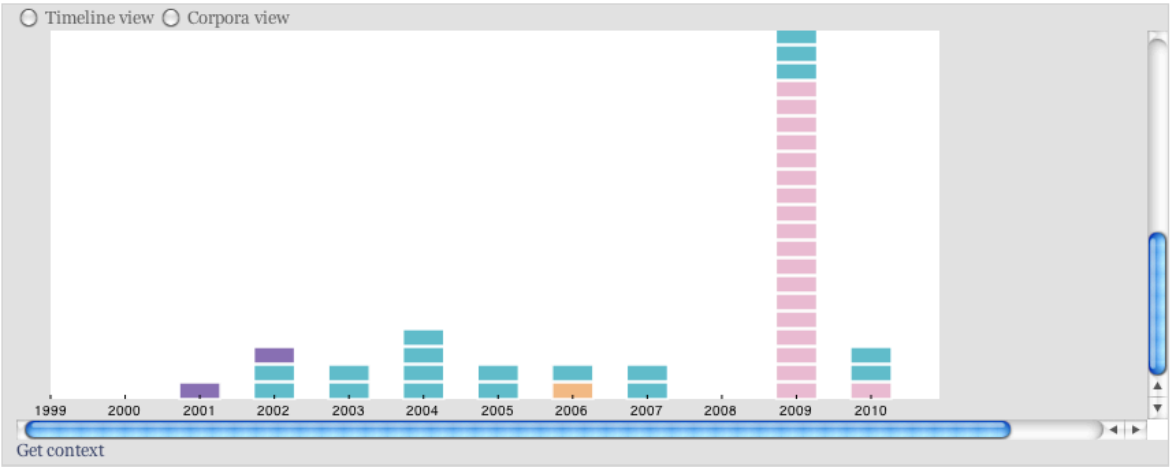
Conscience protection Add to favorite

Found in: 53 documents | Present in the input text



Elective abortions Add to favorite

Found in: 62 documents | Present in the input text



Additional Artifacts