Project Summary

Intellectual Merit

Despite the important role media attention plays in our understanding of politics and society, no existing dataset offers a comprehensive look at how media attention changes over time or whether our understanding of "the" media agenda depends on which news outlet(s) we use to measure it. We propose a two-part study.

First, we will develop large-scale datasets of *New York Times* (1947–2010) and *Washington Post* (1977–2010) news coverage, including *all* front-page stories as well as large (1 in 10) samples of the full set of articles, each coded according to the primary policy or non-policy topic being discussed. We can accomplish this task because of our experience in developing computer assisted content analysis techniques, based on previous NSF support through the Policy Agendas Project and other sources. The resulting datasets, consisting of some 1.2 million news stories classified by topic, will enable researchers for the first time to examine with confidence levels of media attention paid to any topic of interest over several decades.

Second, using electronic searches of 30 U.S. national newspapers, we will develop a dataset of media attention to a sample of over 100 policy issues as well as some non-policy issues from 1985 to 2010. We will also trace attention to these 100+ issues across five television stations from 1998 to 2010, as well as statically on internet news sites and blogs. These data will answer unresolved questions about the coherence or possibly idiosyncratic nature of the topics of coverage of individual news sources in America. Surprisingly, no studies have assessed in the scale we envision the degree to which various media outlets show similar or different levels of interest in policy-relevant issues over time (or to non-policy topics such as sports, entertainment, or the weather, for that matter). As a result, political scientists, sociologists, and others who use media attention as an indicator of salience are dogged by questions of whether their indicators, often based on the *Times*, the *Post*, or the *Readers' Guide*, are accurate reflections of national trends. Our pilot tests suggest that results will justify the Policy Agendas Project's focus on extensive use of just one (or now two) newspapers and that a multiplication of research effort to cover more and more news sources would have little impact on measures of trends over time.

Broader Impact

Scholars in many fields are interested in the dynamics of media attention. The creation of these databases and their free distribution through the Policy Agendas web site (www.policyagendas.org) will facilitate a wide range of studies. The NSF-supported Policy Agendas Project already makes available information about government activities since 1947. The Agendas Project web site has become the standard source for much information about federal government actions, and it is widely used for both teaching and research. However, its *New York Times* database, consisting of approximately 44,000 observations, is too small to support detailed analyses in many policy topics. Further, it is based on a sampling system that relies on the published annual *Index*, which will be discontinued. Improving the quality of this database will enhance a widely used bit of infrastructure, benefitting the profession as a whole.

In addition to these data, we will develop and further enhance computer assisted content analysis techniques and make these algorithms and training applications available through the Policy Agendas web site. These can be used for any large-scale content classification project.

Project Description

Much of politics hinges on attention. At every level of government and in the public realm, the political process is shaped by which issues are discussed and which are ignored. Media attention affects other institutional agendas, public opinion, and public policy (e.g., Baumgartner et al. 2008, Flemming et al. 1999, Iyengar 1991, Jones and Baumgartner 2005, Wood and Peake 1998). Whether media attention is driven purely by events or by events as they interact with institutional and social factors (Bennett et al. 2007, Gans 2004), in order to understand politics we must understand media attention—how it gets distributed across policy areas, and how this attention changes over time.

As in all areas of social science, our understanding of media attention can only be as good as our measures of it. And to date, our measures are quite limited. Certainly the largest media dataset available over a long period of time is that collected by Baumgartner and Jones as part of the Policy Agendas Project, containing 44,246 stories sampled at random from the *New York Times Index* between 1946 and 2004, coded by topic according to the now widely-employed codebook they developed for that project. This dataset represents an enormous data collection effort, and it has provided many key insights into both the media agenda-setting process and the role that media attention plays in shaping the agendas of other institutions (Jones and Baumgartner 2005, Jones et al. 2003, Jones et al. 1997).

Unfortunately, as large as the Policy Agendas Project media dataset is, it is too small by a long shot to provide a concrete understanding of media attention dynamics for most policy topics. Some simple math tells the story easily. Roughly speaking, 44,000 stories over 58 years implies an average of about 750 stories per year. Typically, only 40 to 60 percent of the stories are related to public policy with the remainder being sports, obituaries, weather, book reviews, and other non-policy related content. This leaves an average of 375 policy-related stories per year; just over one per day. Dividing those up by the 21 major topics of the Agendas Project, one sees fewer than 20 stories per topic per year. But the problem is exacerbated by the fact that attention is, of course, not randomly spread across the 21 topic areas: Defense, Crime, International Affairs, and Government Operations (e.g., elections) each have over 10 percent of the total coverage, whereas the areas of Agriculture, Energy, Environment, Trade, Public Lands, Social Welfare, and Science / Technology all have less than three percent coverage. This means that in a typical year the typical number of stories in these areas is in the low single digits (and there are some empty cells in the matrix of stories by major topic by year). Thus, our estimates of media attention are not very robust, and it is likely for this reason that the *New York Times* dataset appears to be one of the least widely used of all the resources distributed through the Policy Agendas website.

Importantly, an additional challenge to using the Policy Agendas Project *Times* dataset is that it is collected by sampling each year of the published *New York Times Index*, which is scheduled to be discontinued.

In sum, there is trouble in Denmark as regards the existing *New York Times* dataset. Fortunately, we have a solution. For her NSF-supported dissertation, Boydstun (2008) compiled a full census of all front-page *New York Times* stories between 1996 and 2006 (more than 30,000 stories, an average of about 3,000 per year), each coded by topic using the same Policy Agendas Project codes. The Boydstun dataset offers meticulous insight into the *Times* front-page agenda, how it operates, and how it changes over time. And it serves to illustrate that collecting a massive media dataset of this size is feasible in a relatively short amount of time. With the help

of undergraduate research assistants funded through an NSF Dissertation Improvement Grant, it took under two years to collect and manually code all 30,000+ stories. The Boydstun front-page dataset is about three-quarters as large as the entire *Times* dataset available through the Policy Agendas site, but because it was collected using more efficient tools—in particular, automated methods of downloading—it was done for a fraction of the cost and in a fraction of the time. In addition, because now we have developed powerful tools of computer-assisted coding, we believe we can code hundreds of thousands of observations using the Agendas Project topic codes quite manageably. (Boydstun did not use computer-assisted coding for her thesis.)

Boydstun's dissertation project also illustrated some important differences between the topics and the dynamics of attention on the front page of the *Times* as compared to the full newspaper. Wolfe, Boydstun, and Baumgartner (2009) compared her front-page dataset with the full *Times* dataset from the Agendas web site as well as the small sample of front-page articles included there. The Boydstun dataset serves to replicate but expands substantially on the sample of front-page stories; the only differences are due to sampling. But the front-page dataset differs in important ways from the content behind Page A1. First, the front-page agenda is more highly concentrated in only a few topic areas, such as elections, defense, and international affairs (e.g., the wars in Iraq and Afghanistan). Second, it shows considerably higher inertia. That is, coverage is "sticky" in that when front-page attention focuses on, say, health care, it tends to stay focused on that, shifting only eventually to the next "hot" issue (see also Boydstun 2008). Inside-page coverage is both more spread out across topic areas and shifts more smoothly from issue to issue as social trends emerge. In terms of policy impact, front-page coverage is a different animal than inside-page coverage, as it rivets attention of policy makers to a select few high agenda items. Thus, for the first and largest component of our study, we propose to develop new databases consisting of the full census of the front page as well as a large-N random sample of the full *Times* and *Post* papers for the period studied. Because of our automated techniques of data collection and computer-assisted coding, this can be done with little additional expense.[1]

The second, but considerably smaller, element of our proposal is a straightforward but large-scale comparison of the coverage of more than 100 policy topics across 30 different newspapers, five different television stations, and (statically) a sampling of internet news sites and blogs in order to assess the reliability of any single news source to reflect the larger national media agenda. Scholars are essentially swimming in the dark as regards their confidence that a single media indicator is enough to capture national trends. Woolley (2000) criticizes scholars for relying on media reports and urges caution in relying on media indices. Soroka (2002), on the other hand, found high levels of correspondence among various Canadian newspapers in their coverage of national policy debates. Baumgartner and Jones (2009) showed reassuring comparisons between the *Readers' Guide* and the *New York Times* across several issues. Baumgartner, De Boef, and Boydstun (2008) found very high correspondences both in the amount of coverage and the type and tone of coverage given to capital punishment across eight newspapers, including papers expected to differ in their treatment of the death penalty, such as the *New York Times* and the *Houston Chronicle*. However, we simply do not have enough information to discern whether we can treat the media agenda as a single one or whether different newspaper, television or internet outlets follow idiosyncratic courses over time, related perhaps to local events rather than national trends. We therefore plan a simple but large-scale

[1] Copyright concerns are the main reason we do not propose a full census of all stories in these papers, as we must download the stories in order to code them. However, since the large 1 in 10 sample we will draw will support fine-grained analysis, we see little added value in collecting all stories for the purpose of discerning trends over time.

comparison of 30 newspapers (including national papers as well as small local papers) across more than 100 different policy-relevant and non-policy issues identifiable with Boolean searches and covering a broad range of topics. This part of our project is not comprehensive, as the first part is (i.e., our set of 100+ issues is not exhaustive), but we hope that it will be large enough to answer definitively some of these questions about the correspondence or discrepancy between news sources (and, according to our preliminary results, to put some of the concerns to rest).

We focus on newspapers rather than, for instance, the recently blossoming sources of information on the internet because we want to cover a long time period and because newspaper coverage is most relevant to the first part of our study, where we collect and code *New York Times* and *Washington Post* articles. However, the potential exists of course that traditional newspapers are similar to each other but differ systematically from other types of information sources, including television news programs, internet news sites, and especially blogs. Thus, this second part of our study will include tracing attention to our 100+ policy and non-policy issues across five television stations (ABC, CBS, NBC, CNN, and FOX) between 1998 and 2010, using the same Boolean search terms applied to electronic news transcripts. And while we cannot study internet coverage in a dynamic manner, we will also include a static comparison of levels of coverage to the same 100+ issues in a sample of electronic internet news sites and news blogs.

Combined, the two parts of this project promise to answer pressing questions in political science about the role of media in politics. And beyond the significance our study holds for political science, we expect that scholars in many other fields will benefit from the resulting datasets, as well as the subsequent scholarship we intend to develop based on our findings. Sociologists could use the data to understand how the congestion of real-world events pushes some social movements out of the news while giving disproportionate attention to others. Communications scholars should be interested in both datasets we will produce. Historians may find the 64-year time series of media attention of interest, especially since it can be directly compared to government outcomes. In short, these and other fields that rely on measures of news coverage to examine the behavior of media attention both as a dependent and as an independent variable need much richer datasets in order to perform robust investigations. Finally, computer scientists and others concerned with automated classification systems may find the large, text-based, and high quality databases we will create to be of interest as they attempt to refine and develop more accurate and efficient text classifiers, especially since many different types of text-based databases are available in addition to the one we propose to create here. The Agendas Project is into an important tool for a variety of user communities.

PART 1: *New York Times* and *Washington Post* Datasets

As described above, the first part of our study involves all front-page articles and a one-in-ten sample of the entire newspaper for the *New York Times* (1947-2010) and the *Washington Post* (1977-2010).[2] Rough estimates of the expected numbers of observations are as follows:

1. All *New York Times* front-page stories, 1947–2010 (N≈175,000)
2. A 10% sample of all *New York Times* stories, 1947–2010 (N≈700,000)
3. All *Washington Post* front-page stories, 1977–2010 (N≈80,000)
4. A 10% sample of all *Washington Post* stories, 1977–2010 (N≈225,000)

[2] LexisNexis offers electronic versions of news stories for the *Post* back to 1977 and the *Times* back to 1985. For the 1947 to 1984 *Times* data, we will employ Abby optical character recognition software to transform into text form the articles archived in PDF format in ProQuest's *New York Times* Historical Archive.

What makes us think we can create these databases, consisting of 1.2 million stories, in the span of two years, when the current media database of the Agendas Project has less than 50,000 observations? Many developments point to the surprising feasibility of this endeavor, mostly related to skills we have developed through other elements of the Policy Agendas Project. Indeed, several databases associated with the project, such as the Congressional Bills data, number in the hundreds of thousands of observations. But two elements are key. First, we can *automatically download* the raw data using web scrapers. This process is straightforward now and so we do not explain it in detail. More important is the development of *computer assisted coding* techniques we have developed with computer scientist Paul Wolfgang at Temple University. Wolfgang has worked with Baumgartner through the Pennsylvania Policy Database, the first project at the state level designed to replicate the national-level Agendas Project. Working to enhance the efficiency of the NSF-supported Congressional Bills Project (http://congressionalbills.org), Stephen Purpura and John Wilkerson developed supervised learning algorithms to code hundreds of thousands of bills into the four-digit Policy Agendas codes (see Hillard, Purpura, and Wilkerson 2007). Considering the large amount of text-based data already coded to high standards of reliability through the Agendas Project, it is a natural for those interested in supervised learning techniques to apply their tools to our data. Purpura first developed these tools and showed impressive results (e.g., 89 percent accuracy at the two-digit level and 81 percent at the four-digit level), using very large training datasets in the Bills Project. In the Pennsylvania Policy Database Project (http://www.temple.edu/papolicy) Wolfgang was central to the effort to apply these tools to the specific problem of coding abstracted newspaper clippings as well as various legislative texts (e.g., bills, laws). Working with Penn State graduate student Jon Moody, he developed tools allowing us to code tens of thousands of clippings based on a sample of hand-coded "training" data. The automated coding (or computer assisted coding, as we call it here) techniques are now being routinely implemented across the Comparative Policy Agendas Projects and are working in different languages and on different types of documents, ranging from newspaper stories to legislative documents to executive speeches (see http://www.comparativeagendas.org/text-tools). Moody and Wolfgang are consultants on this proposal, and we expect further development and refinement of the tools as we go forward.

Computer Assisted Coding

Many political scientists are now working to develop supervised or unsupervised text annotation systems for a variety of purposes (see, e.g., Klebanov et al. 2008, Hopkins and King n.d., Landauer and Dumais 1997, Laver et al. 2003, Lowe 2004 & 2008, Monroe and Maeda 2004, Monroe et al. 2008, Simon and Xenos 2004). Our purposes are quite simple: We want to replicate exactly the Policy Agendas classification system, and to do so with very high accuracy. Fortunately, this goal is entirely possible. Further, it does not require the development of any new techniques. Rather, it requires the active collaboration of substantively focused political scientists (such as those involved with creating the Policy Agendas Project) with those familiar with the computer science applications already in existence.

Different automated content analysis programs work in different ways, but most operate on two key principles: each word in a text can be treated as a unit of data, and documents (speeches, newspaper articles, legislative bills) can be characterized by the relative occurrence of various words. For example, a story that uses the words Detroit Tigers, hit, run, ball, bat, and homerun, but no other specialized vocabulary, can clearly be recognized as a baseball story even if it does not contain the word baseball. Every text can be uniquely identified by its word

frequencies, relative to some baseline. Most systems use "bag of words" approaches, in which the program ignores syntax and word combinations, though we also incorporate some "n-gram" approaches, which use combinations of words occurring in sequence. More complicated systems that incorporate sentence structure, parts of speech, syntax, and other factors exist but are not necessary for our purposes. Key to our system is to use several different algorithms and to take cases where all predictors agree; this approach produces the most accurate results. However, it also reduces coverage, meaning that it reduces the number of cases where multiple algorithms classify a text in the same way. If different algorithms produce disparate predictions for a given text observation, then that observation is not "covered." Multiple iterations, successive hand-coding of samples of the uncovered cases, and the use of large previously coded validation datasets allow us to maximize coverage while maintaining high predictive accuracy.

The specific set of programs we employ for this project is called TextTools, a text classification program developed by Paul Wolfgang. The TextTools program consists of five separate algorithms, each designed to take a slightly different approach to the same problem: Given a collection of texts that have already been coded by topic area, assign topic codes to ("classify") a collection of texts that have not been coded. In the language of automatic text classification, the texts that have already been coded, or classified, are called "reference" texts, while the un-coded texts that we want to classify are called "virgin" texts. A researcher trains the TextTools program by importing a corpus of previously coded reference texts and their corresponding codes. During the learning phase, the algorithms calculate the distinctive patterns of word usage that distinguish those texts coded into the different categories. In the prediction phase, it applies these models to each virgin text and assigns ("predicts") the most likely category for that text. Testing the accuracy of this process is simple enough: One just trains the model on a portion, rather than all, of the previously coded reference texts, and has the computer "predict" the codes that humans have already assigned. Comparison is then simple.

There's nothing magical about this process of identifying the content of texts we have never read. The TextTools algorithms provide speed, computing muscle, and consistency, but the basic process of using words as data to classify a text by topic is very straightforward. Consider, for example, the following three paragraphs from a recent story entitled "New Medicare Plan For Drug Benefits Prohibits Insurance:"

> *"Medicare beneficiaries will not be allowed to buy insurance to cover their share of prescription drug costs under the new Medicare bill to be signed on Monday by President Bush, the legislation says.*
>
> *"Millions of Medicare beneficiaries have bought private insurance to fill gaps in Medicare. But a little-noticed provision of the legislation prohibits the sale of any Medigap policy that would help pay drug costs after Jan. 1, 2006, when the new Medicare drug benefit becomes available.*
>
> *"This is one of many surprises awaiting beneficiaries, who will find big gaps in the drug benefit and might want private insurance to plug the holes -- just as they buy insurance to supplement Medicare coverage of doctors' services and hospital care"* (Pear 2003).

Table 1 presents a "bag of words" presentation of this content, showing the frequency of each root word and its variants. Whether we read the article as a human would or look at the frequency of the words employed, which is analogous to what the computer algorithms do, we easily conclude that the story is about legislation concerning Medicare prescription drugs, which

is all we need to know in order to code the story into a Policy Agendas Project 4-digit topic code. Certainly words such as "Medicare," "drug," and "insurance" could appear in stories about any other topic as well (in stories about the economy, social security, or employment benefits, for example), but given the relative high frequencies of their use in these paragraphs and given the absence of any other high-frequency words indicative of an alternate topic, all five of the TextTools algorithms are able accurately to classify this text.

Table 1. Word Frequencies for an Article on Health Care

Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq
the	7	new	2	Bush	1	Medigap	1	says	1
Medicare	6	private	2	but	1	might	1	services	1
of	6	will	2	by	1	millions	1	share	1
to	6	1	1	care	1	Monday	1	signed	1
drug	4	2006	1	doctors	1	not	1	supplement	1
insurance	4	a	1	fill	1	on	1	surprises	1
beneficiaries	3	after	1	find	1	one	1	that	1
and	2	allowed	1	have	1	pay	1	their	1
be	2	any	1	help	1	plug	1	they	1
benefit	2	as	1	holes	1	policy	1	this	1
buy	2	available	1	hospital	1	prescription	1	under	1
costs	2	awaiting	1	is	1	President	1	want	1
cover(age)	2	becomes	1	Jan.	1	prohibits	1	when	1
gaps	2	big	1	just	1	provision	1	who	1
in	2	bill	1	little-noticed	1	sale	1	would	1
legislation	2	bought	1	many	1				

Multiple Iterations Between Hand and Machine Coding

Our use of automated classifiers for newspaper stories is greatly facilitated by the fact that we already have very large training datasets of electronic full-text news stories available. Boydstun has coded over 30,000 *Times* stories already, and we have done some preliminary work to see if there are important differences between the *Times* and the *Post*. (If the two tend to be linguistically distinct, we could not use data from one to code reliably data from the other, but they appear to be quite similar.) We also have experience from other applications of TextTools, such as the Pennsylvania newspaper clippings project.[3] The steps for testing computer-assisted classification are as follows, and below we present some indicative results.

i. Train using a large reference dataset and a 90/10 iteration[4]
ii. Apply the algorithms to virgin texts that have not been coded
    a. Identify topic areas with > 90 percent accuracy and take resulting data into training dataset for the next round

[3] As part of the Pennsylvania policy database, we used TextTools to code abstracts our students wrote from Xeroxed copies of the articles included in the Governor's daily news briefings.

[4] We follow this procedure. Assume 10,000 observations in the reference dataset. Select 1,000 at random. Use the other 9,000 to train the algorithms, and predict the 1,000 randomly selected ones, calculating the percentage accuracy for each topic area. Repeat this process 10 times so that each case is rotated through being part of the training and virgin data. Take the average of the 10 calculated accuracies. (This procedure allows us to use 90 percent of the data to predict 10 percent, rather than the simpler 50/50 breakdown, and this approach has proven to be the most efficient in previous experience.) In the end, it allows us to know the topic areas (typically those with the most observations) where accuracies are very high and those where accuracies are below our standards.

b. Iterate with the growing training dataset until there is no further benefit
c. Code by hand random samples of virgin texts in topic areas we know from our pilot study to be poorly predicted, and incorporate larger numbers of these into the training dataset, iterating after every 1,000 newly coded cases

Result: Only a minority of cases must be coded, such that human energy can be reserved for the "hard" cases. Of course, the logistics of this approach only work if the classifiers achieve high results once the training datasets are large enough. We present encouraging results below.

Vocabulary Drift and Distance between the Training and Predicted Datasets

We must be sensitive to the problem of "vocabulary drift" considering the long historical period we seek to cover. Individual proper names and names of such agencies as the "Environmental Protection Agency" are very useful for the classifier algorithms, as these are often highly discriminating, accurately suggesting that the story in question relates to the environment. But we can't use a classifier trained on data from the 1990s or 2000s to classify data from the 1950s or 1960s. Experience from the Congressional Bills Project suggests the solution: only use data from contiguous years, since accuracy declines as the time difference increases. We will have to experiment in order to determine the proper time periods, which might be years or slightly longer periods. In any case, we will work backwards in time, classifying large numbers of cases based on training datasets from the same historical period, then using the cleaned and accurate database from one period as the training dataset to classify the previous historical period, moving in steps so as never to be moving in historical period by more than one or a few years at a time.

Training TextTools on one type of document, but asking it to predict the values of another may or may not produce good results depending on the linguistic characteristics of the documents. For example, congressional bills would predict laws and possibly hearings, but may not do well for newspaper stories, as the type of language used can be highly different, and many newspaper stories are on topics that are never or only rarely the object of any attention in a legislative setting (e.g., sports results, routine business news, restaurant reviews). We expect that the *Washington Post* and the *New York Times* will be relatively similar in their vocabularies, and preliminary analysis confirms this assumption.[5]

Pilot Study Results

We present the results of our pilot study here, focusing on the implications of feasibility for our larger proposed study. Table 2 shows the results of our first pass at a 90/10 iteration, using the 30,780 records in Boydstun's hand-coded *Times* front-page dataset as both the training data and the data to be predicted. Successively through ten iterations, 90 percent of the observations were used to train three different algorithms to predict the remaining 10 percent of the data. The table presents the average results of ten iterations of this process, each one on a different random selection of 10 percent of the dataset. The first three columns show, respectively, the topic category number, the topic name, and the average number of cases actually coded in each topic category. The next series of columns show the accuracy of three individual classifiers, SVM, Lingpipe, and MaxEnt; that is, the percentage of the stories in each category that each algorithm

[5] In a test using Boydstun's *NYT* dataset to predict a small sample of 846 *WP* stories conducted for this proposal, the first iteration of TextTools achieved 76% accuracy, as opposed to 79% for the *NYT* as shown in Table 2. Of course, this accuracy rate will be improved through iteration and larger databases, but as an initial estimate it suggests there may not be that much difficulty in coding this second source based on existing codes for the first. Of course, as the project develops we will have larger reference databases for both papers.

Table 2. Accuracy Tests of Three Automated Classifiers.[6]

			Each of Three Algorithms			Combinations of Two Algorithms						All Three Algorithms	
			I	II	III	I and III		II and III		I and II			
Code	Topic	N	Acc'y	Acc'y	Acc'y	Cov'g	Acc'y	Cov'g	Acc'y	Cov'g	Acc'y	Cov'g	Acc'y
1	Economy	84.8	2.59	40.80	21.70	52.71	3.58	38.21	40.12	31.49	7.12	20.99	8.43
2	Civil Rights	86.3	2.32	26.30	19.81	50.29	3.92	33.26	32.06	30.94	6.37	18.66	9.32
3	Health	182.2	4.99	62.40	12.68	51.15	5.58	22.83	48.80	19.37	24.36	11.36	23.19
4	Agriculture	16.1	0.0	9.94	3.73	51.55	0.0	23.60	7.89	25.47	0.0	13.66	0.0
5	Labor	74.2	4.18	24.93	16.17	47.04	7.45	31.13	30.74	33.42	8.87	19.41	13.19
6	Education	87.5	5.71	44.23	15.31	41.60	11.81	23.20	45.32	18.86	26.06	12.23	35.51
7	Environment	34.2	0.88	12.87	5.26	47.37	1.85	23.10	10.13	29.24	2.00	16.08	3.64
8	Energy	29.4	0.68	26.19	7.82	51.02	1.33	26.53	23.08	26.87	2.53	16.33	4.17
10	Transport	58	5.52	33.10	15.52	45.00	11.11	30.00	43.10	26.21	19.74	16.38	30.53
12	Crime	174.5	8.42	53.35	25.10	43.21	14.72	33.07	59.97	26.07	28.13	15.13	37.50
13	Welfare	26.5	1.13	19.62	22.26	44.53	1.69	33.21	31.82	27.55	4.11	20.00	3.77
14	Housing	40.1	0.0	14.96	11.97	31.92	0.0	22.69	15.38	18.95	0.0	9.23	0.0
15	Commerce	149.2	60.72	60.32	53.89	65.21	75.23	49.46	82.93	54.56	84.64	43.30	**88.39**
16	Defense	376.8	11.25	59.10	51.14	51.30	19.81	54.64	66.15	40.90	23.49	26.46	33.30
17	Science	82.7	4.11	39.78	22.01	47.28	8.18	30.11	47.39	25.39	12.38	14.15	21.37
18	Trade	24.8	0.0	11.29	5.65	63.31	0.0	45.16	6.25	54.03	0.0	37.10	0.0
19	International	705.8	**94.45**	82.54	73.00	75.90	**95.28**	69.68	**92.33**	81.34	**98.07**	65.20	**98.04**
20	Government	388.2	**88.28**	79.39	82.30	82.10	**95.36**	72.00	**96.06**	77.74	**96.06**	68.52	**97.89**
21	Lands	25.3	0.0	11.46	11.86	39.13	0.0	25.69	15.38	26.88	0.0	13.44	0.0
23	Culture	75.8	19.53	58.05	40.63	41.82	41.64	39.71	76.08	26.25	60.30	19.26	75.34
24	State Gov't	74.7	6.16	38.55	26.91	39.09	15.41	34.54	54.26	29.72	18.92	18.07	30.37
26	Weather	55.9	9.12	50.63	31.48	46.15	19.38	43.47	53.50	29.70	28.92	23.97	35.07
27	Fires	12.9	0.0	23.26	8.53	26.36	0.0	17.05	27.27	12.40	0.0	3.88	0.0
29	Sports	128.9	79.60	**89.06**	**88.36**	85.42	**92.73**	84.25	**98.25**	78.59	**97.04**	77.19	**98.39**
30	Obituaries	19.8	0.0	41.41	37.37	27.27	0.0	37.37	51.35	16.16	0.0	11.11	0.0
31	Religion	44.5	0.90	38.43	17.98	46.29	1.94	34.16	34.87	29.44	2.29	17.53	3.85
99	Other	16.9	1.18	11.24	15.38	27.22	4.35	34.32	29.31	14.20	8.33	8.88	13.33
(Total) & Averages		(3,076)	42.70	60.26	48.25	60.43	58.35	51.09	76.96	50.73	71.49	39.17	78.80

[6] For 3,076 observations coded by hand, the cell entries show the percentage accurately predicted by each of three automated classifiers (I = SVM, II = Lingpipe, III = MaxEnt), by combinations of two of them, and by the combination of all three. In the case of combined accuracies, we also report the "Coverage," which is the percent of cases where the different algorithms make the same prediction as one another. So, for example, the last cell of the last row shows that across all of the observations for which all three classifiers predicted the same code, these predictions were accurate 78.8 percent of the time, but the cell directly to the left shows that only 39 percent of all cases fall into this covered category. Numbers in bold show where over 88 percent accuracy or higher was achieved. Sports, government, international, and business news have the highest accuracies. These categories are also based on the largest N's.

accurately predicted as belonging to that category.[7] Then we show the combinations of results where two classifiers provide the same prediction. "Coverage" declines because the two do not always make the same prediction, but "accuracy" invariably increases. Finally, the last two columns show coverage and accuracy where all three predictors agree.

Several things are worth noting here. First, accuracies vary dramatically by topic area. Largely this is because of the number of observations in the training dataset. The computer cannot learn to recognize things it has not seen, and where it has only seen a few examples of a topic its predictive accuracy is low. Moreover, the nature of the *Times* is to be highly skewed by topic, as discussed above. There are many stories on elections, international affairs, sports, and general business news, but relatively few stories, especially on the front page, about agriculture, public lands, and state government activities. For our purposes, this can be a positive, as it means that if we can accurately predict just six topics (sports, international, defense, commerce, crime, and health) we would have two-thirds of our project complete. In the first estimate, accuracies for defense, health, and crime stories are not yet very high. However, commerce, international, and sports stories are recognized by the algorithms with relatively high accuracies.

Second, combinations of algorithms are more accurate than single ones, but we can investigate the loss of coverage as it relates to increased accuracy on a topic-by-topic basis in order to decide on the best trade-off between increased accuracy and decreased coverage.

Third, we can use the results from this first iteration to note areas where there is apparent confusion (for example, looking in detail at the incorrectly predicted stories in defense suggests that many were coded as international), suggesting that we may want to clarify our coding rules and/or provide more observations in the training dataset to make this distinction more clear. Further, in topic areas where there are simply too few observations in the training dataset, we can use the predicted codes as a means of sifting through tens of thousands of observations to create training datasets that have enough observations in each topic area for the computer to learn these accurately. We can use these predicted values, applied on large virgin datasets, to build our training datasets in a more efficient manner, ensuring better spread across topic areas, essentially creating a training dataset with a minimum number of cases in each topic area. Without such an optimizing procedure, we use our human coding resources highly inefficiently as they generate more than enough cases in some topic areas and insufficient numbers in most topic areas.

Fourth and most importantly, the large gains we expect in predictive accuracy over the course of the project stem from an iteration not demonstrated here but which will be central to the process we will follow if we are awarded the grant. With each round of the use of our automated classifiers, we will feed back into the training dataset all those cases coded in topic areas where the accuracy is above approximately 90 percent, equal to the accuracy we get with human coding. And with correcting a sample of cases coded in topics with low accuracies, we will get more and more cases in those topic areas with few cases in the training dataset. So the training dataset will grow larger and larger over time, with entire groups being added to it *en masse* as stories on such topics as elections and international affairs are accurately recognized by our classifiers, and as our students code a few hundred additional observations in those topic areas not yet accurately predicted. In the end, we expect to code only a small percentage of the overall cases but to focus our hand-coding efficiently by multiple iterations of this process.[8]

[7] TextTools includes five classifiers but we use only three for this illustration. As we discuss below, we will investigate empirically the gains in coverage and accuracy we get from each classifier considered alone and in combination.

[8] This process will get extremely demanding in terms of computing processing power and memory as the increasingly large training datasets incorporate tens of thousands of full text documents. This explains our request

Finally, as discussed above, we will use training datasets only from historical periods relatively close in time to those periods we are classifying. This will require additional hand-coding, but we believe we can identify the most prominent sources of errors in this process by constant observation of the patterns in errors predicted by our classifiers. The net result we seek is not an innovation in computer science but simply the intense, routinized, and repeated application of these existing classifiers in order to increase the efficiency of the focus of our student coders. By focusing their efforts carefully, we expect at least an 80 percent increase in efficiency, resulting in a clean but huge database that will increase by several fold our ability to measure and examine media attention.

PART 2: Cross-News Outlet Comparison

For the second part of our study, we identify 100 policy issues, selected to ensure coverage of each of the 21 major policy topics in the Agendas Project system, as well as a number of non-policy related items. These 100+ issues will not necessarily correspond perfectly to the 4 digit sub-topics of the Agendas Project, as these cannot always be recreated through simple Boolean searches. However, our goal will be to produce a set of accurate Boolean searches that cover the full range of American public policy and to compare these searches, as well as searches designed to track such non-policy topics as sports, entertainment, and weather, so we can know if various news outlets follow similar patterns over time; if there are any systematic differences in the types of topics where there is higher or lower correspondence; and if some news sources follow national trends more or less than others. Specifically, we will trace attention to the 100+ issues across 30 different newspapers from 1985 to 2010, including both large and small papers as illustrated in our pilot study below; news transcripts from five television stations (ABC, CBS, NBC, CNN, and FOX) from 1998 to 2010; and a static sampling of internet news sites (e.g., nytimes.com, msnbc.msn.com) and blogs (e.g., The Huffington Post, Instapundit).

Pilot Study Results

To test the viability of this second part of our study, we developed simple keyword searches to identify news stories on six policy-related and non-policy-related topics: snow storms, Medicare, inflation, the Boston Red Sox, gubernatorial elections, and Brad Pitt. We repeated each search several times, once for each month of the time period under investigation (here, 2001–2003), and we conducted each search across several newspaper outlets. We focus here on how our results compare across a selection of twelve different newspapers, both national and local in focus: Boston Globe, Boston Herald, Charleston Gazette (WV), Chicago Sun-Times, The Columbian (WA), Houston Chronicle, New York Times, Philadelphia Inquirer, Pittsburgh Post-Gazette, St. Paul Pioneer Press (MN), St. Petersburg Times (FL), and Washington Post.

The question we're concerned with is not whether the absolute levels of coverage for each topic are similar across newspapers, but whether the amounts of coverage given to a topic

---

for several large capacity computers. Our pilot study revealed that standard computers with dual-core processors were incapable of running the TextTools software on such large text datasets; the program routinely crashes on these standard machines. And beyond the fundamental need to have enough computing power to run TextTools in the first place, we estimate that we need four high-end desktop computers, two on each campus, in order to complete our data collection and analysis in the proposed two-year time frame. We expect these computers to be tied up for hours at a time during the process of article retrieval and during each of the multiple iterations that we envision.

by different papers change in similar ways over time. For example, even if *The Columbian* runs one story for every ten *Houston Chronicle* stories about inflation, what matters is whether inflation rises and falls on both agendas in the same way over time. Thus, we employ factor analysis to compare the dynamic patterns in coverage of each topic across papers. By subjecting counts of the number of news stories published on the topic at hand each month in each newspaper to principal-components factor analysis, we can indentify statistically the extent to which these different newspapers "hang together" over time.

Table 3. Factor Loadings on Twelve Newspapers for Six Different Topics.[9]

Newspaper	Snow Storms (N=1,333)	Medicare (N=12,801)	Inflation (N=15,703)	Boston Red Sox (N=29,714)	Gubernatorial Elections (N=16,006)	Brad Pitt (N=1,497)
Boston Globe	**0.91**	**0.90**	-0.03	**0.87**	**0.86**	**0.89**
Boston Herald	**0.90**	0.79	0.19	**0.83**	**0.89**	0.53
Charleston Gazette (WV)	**0.83**	**0.85**	**0.87**	0.11	-0.08	-0.07
Chicago Sun-Times	0.76	**0.85**	0.72	0.62	0.70	0.22
The Columbian (Vancouver, WA)	**0.85**	**0.91**	**0.92**	0.72	0.22	0.01
Houston Chronicle	**0.85**	**0.94**	**0.89**	**0.87**	0.55	0.55
New York Times	**0.93**	**0.90**	**0.90**	**0.85**	0.75	0.65
Philadelphia Inquirer	**0.94**	0.68	**0.84**	0.74	0.58	0.44
Pittsburgh Post-Gazette	**0.87**	**0.86**	**0.88**	0.65	0.62	0.56
St. Paul Pioneer Press (MN)	0.78	**0.89**	**0.87**	**0.86**	**0.84**	0.78
St. Petersburg Times (FL)	**0.80**	0.58	0.46	**0.80**	**0.85**	0.62
Washington Post	**0.93**	**0.94**	**0.90**	**0.80**	**0.84**	0.47
Averages	**0.86**	**0.84**	0.70	0.73	0.64	0.47
% Variance Explained by First Factor	74%	72%	59%	57%	48%	30%

Table 3 shows the rotated factor loadings produced from each factor analysis of topic coverage for each newspaper. These results suggest that "the media" really does operate as a central entity, with most news outlets attending to the rise and fall of real-world issues more or less in tandem. The newspapers featured in Table 3 load at a level of 0.80 or better in well over half the observations and, as shown in the final row, the first factor explains nearly half the variance or more in each case, Brad Pitt being an exception. These results tell us that the factors represent dominant patterns of attention dynamics governing news coverage across the country.

[9] The cell entries show the rotated loadings for the first factor acquired by performing principal-component factor analysis on each of six different datasets, each one containing the number of news stories printed on the given topic in each newspaper by month. These counts were obtained by running monthly searches in LexisNexis for the following keywords: for snow storms, <snow AND blizzard>; for Medicare, <Medicare>; for the Boston Red Sox, <"Red Sox" AND NOT cap>; for inflation, <inflation! AND price! AND (United States OR U.S. OR US)>; for gubernatorial elections, <governor! w/5 (elect! OR vote!)>; for Brad Pitt, < "Brad Pitt">. Data extends from 2001 to 2003, except for inflation (2001–2006). At the top of the table, the total number of stories retrieved from these twelve newspapers for each topic across the entire time period is shown in parentheses. Throughout the table, numbers in bold show those factor loadings that are 0.80 or higher, indicating strong commonality with the latent dynamic trend identified by the first factor. The final row in the table shows the percentage of the total variance in each dataset that is explained by the first factor; in other words, how strong the most dominant latent dynamic trend is across the newspapers in the dataset. Note that this preliminary test is based on just 36 monthly observations.

Figure 1. Coverage of Medicare Across Five Major Newspapers, 2001–2003.

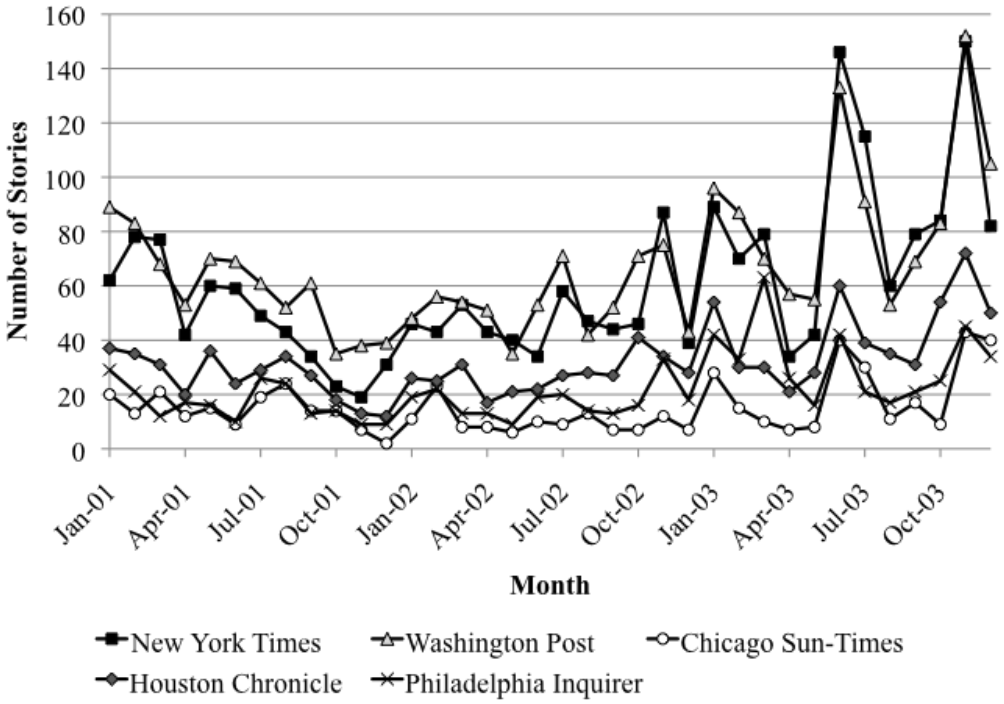
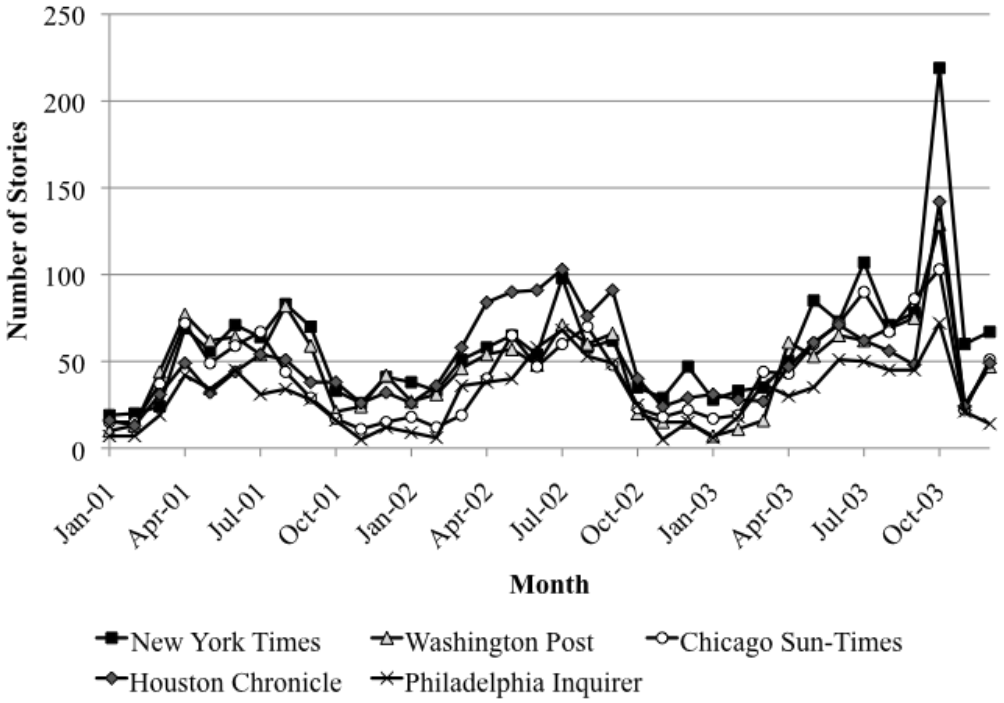


Figure 2. Coverage of the Boston Red Sox Across Five Major Newspapers, 2001–2003. [10]



[10] The two Boston papers, of course, have much higher *levels* of coverage of their hometown baseball team. Table 3 shows, however, that the trends over time are very similar, as they load very highly on the first factor based on all twelve papers.

Importantly, not all newspapers we might examine trend as closely together; the *Wall Street Journal*, for example, seems to march to the beat of its own drum, generally loading only at low levels when included in our factor analyses. The lesson there is that some news outlets simply are not representative of "the media" agenda, and studies interested in using one or more news outlets as proxies for national attention need be wary of employing those sources for that purpose. Our proposed study would include detailed analysis of at least 30 different newspapers in their coverage of 100 different sub-topics, allowing us to separate these wild card sources from the rest of the pack.

However, the main story from this pilot study is quite clear: Most U.S. newspapers, most of the time, reinforce a common national media agenda. This conclusion is illustrated in Figures 1 and 2, which offer "eyeball" tests of how well five of the largest newspapers in our pilot study hang together over time. Whether the topic is one of national importance, like Medicare, or a matter of regional culture, like the Red Sox, news coverage to these issues tends to rise and fall in tandem across newspapers.

Dissemination

All data collected as part of this project will be integrated into the datasets distributed through the Policy Agendas Project website. Data from Part One of this proposal will be integrated into the "policy analysis tool" which allows users to conduct "on the fly" analyses or to download any subset of the data; for example, all coverage of a certain policy topic, or data only for certain years. It will be integrated with the other data resources of the Agendas Project as well (e.g., congressional hearings, presidential speeches, etc.), allowing one to request both media and government attention to any policy topic. Because of copyright concerns, we will not distribute the raw text files associated with the news stories. However, we do expect to produce the date, headline, and a URL through which a user can link to the LexisNexis or ProQuest site through their university library to view the full text of the stories they select. The Pennsylvania Policy Database project is implementing such an internet tool for Pennsylvania Supreme Court decisions available only through subscription with Westlaw, Inc. We expect to follow this model for the newspaper stories. Of course, users who want only to know the number of stories on a given topic per year will be able to see those results immediately.

Data collected in Part Two of this project will also be available through the same web site, but these will not be integrated into the policy analysis tool, as they will not correspond to the topic and subtopic codebook which integrates the various databases of the Agendas Project. Rather, each individual search will be made available in a separate file showing the number of hits per month in each newspaper and television station, as well as the static counts for attention to each issue on the sampled internet sites and blogs.

The dissemination of data for the widest possible scholarly and educational use is the essence of the Policy Agendas Project, and these data will constitute an important element of that project. Therefore we will be particularly attuned to the dissemination aspects of the data we collect. Michelle Wolfe, consultant on this project, is project manager for the Policy Agendas Project and will provide continuing support not only for training our students in the details of the content coding system, but also for providing liaison with the programming and website development team at the University of Texas at Austin as they redesign the web site to incorporate additional datasets.

Conclusions

Systematic analysis of the links between media attention and government activity requires higher quality measures of media coverage by policy topic than currently exist. The Policy Agendas Project has put in place virtually all the tools to make possible such analyses, but its existing media datasets are too small to support detailed analysis for most individual policy topics. Advances in internet retrieval mechanisms and computer science text-analysis research now make possible the collection of datasets larger by orders of magnitude than previously existing ones. We therefore can take off-the-shelf computer science applications and, with considerable man-machine interaction, iterate through a process that will result in more than a million newspaper stories from two major newspapers, one covering the entire post-World War II period, so that we can finally achieve for the media what the Agendas Project has already done for many aspects of government: the creation of highly accurate indicators of trends in attention over time for the full gamut of public policy concerns. Our data collection project goes beyond only public policy concerns and, therefore, will be of interest to an audience well beyond only political science. We expect students and scholars in many disciplines to make wide and frequent use of the data we collect.

The second part of our project includes the comparison of 30 newspapers and five television outlets (as well as internet news sites and blogs), allowing a large-scale assessment of the degree to which attention to various aspects of public policy (and non-policy related stories) is idiosyncratic, with individual news sources independently following agendas driven by editorial discretion or local events, or systematically associated with a single national trend over time. Of course, we expect variance in both directions: Some newspapers, for example, may be more fully integrated into a national agenda while others may be more locally focused, and some topics may show greater national coherence in coverage than others. Detailed empirical knowledge of the dynamics of these questions is almost completely lacking. Yet it is needed in order to assess the quality of existing estimates of media attention to public policy concerns.

A third outcome of our project is the refinement of the TextTools software, including the creation of a user-friendly interface allowing non-specially trained social scientists to conduct their own computer-assisted coding projects. This web interface would increase the value of the project by making available the tools we develop to an audience that may not be able to collaborate actively with professional computer scientists. In our experience, this has severely restricted the use of readily available, but unfamiliar tools in the social science community.[11]

In sum, we propose new data resources previously not possible to be created and distributed to the widest possible community. Such a contribution to the infrastructure of social science research will allow a wide variety of empirical and theoretically driven projects to be conducted. We will conduct our own theoretically driven research of course, focusing especially on the determinants of front-page newspaper coverage and the impact of such highly salient attention on government agendas. However, others in the broad social science community may well have other concerns, and our focus here is on the creation of broadly applicable and high quality data resources. The creation of these datasets requires the use of currently available but poorly understood tools from computer science, such that another output of our project will be the wider dissemination of these highly useful tools to the broader social science community.

[11] This was one of the major conclusions of the NSF-supported conference convened by Baumgartner and John McCarthy in 2007 bringing together computer scientists and social scientists involved in large scale data collection projects.

Results from Prior NSF Support

Boydstun: No prior NSF support
Baumgartner (not including dissertation grants beyond Boydstun's or REU supplements):
SES 0617492, July 1, 2006 to June 30, 2007. Dissertation award for Amber Boydstun, "Doctoral Dissertation Research in Political Science: Agenda Setting and Issue Framing Dynamics on Front Page News."

- Successfully completed dissertation (Boydstun 2008). Book manuscript to be submitted to Cambridge University Press, fall 2009. Forms the basis for this proposal.

SES 0719703, September 1, 2007 to August 31, 2008. "New Computer Science Applications in Automated Text Identification and Classification for the Social Sciences."

- This has led to the development of automated classifiers for the Policy Agendas Project to be included in this research project.

SBR 0111611, January 1, 2002 to December 31, 2007. "Collaborative Research: Database Development for the Study of Public Policy."
SBR 9320922, March 15, 1994 to February 28, 1998 "Policy Agendas in the United States since 1945."
These two awards have supported the Policy Agendas Project. Results include:

- www.policyagendas.org
- www.comparativeagendas.org
- *Agendas and Instability in American Politics*, 2nd ed. Chicago: University of Chicago Press, 2009 (with Bryan D. Jones).
- A General Empirical Law for Public Budgets: A Comparative Analysis. *American Journal of Political Science*, October 2009. (Bryan D. Jones, Frank R. Baumgartner, Christian Breunig, Christopher Wlezien, Stuart Soroka, Martial Foucault, Abel François, Christoffer Green-Pedersen, Peter John, Chris Koske, Peter B. Mortensen, Frédéric Varone, and Stefaan Walgrave)
- Punctuated Equilibrium in Comparative Perspective. *American Journal of Political Science*, 53, 3, (July 2009): 602–619. (Frank R. Baumgartner, Christian Breunig, Christoffer Green-Pedersen, Bryan D. Jones, Peter B. Mortensen, Michiel Neytemans, and Stefaan Walgrave)
- *Comparative Studies of Policy Agendas*. New York: Routledge, 2008. (Edited, with Christoffer Green-Pedersen and Bryan D. Jones). (Previously published as a special issue of the *Journal of European Public Policy*, vol. 13, no. 7, September 2006.)
- *The Politics of Attention: How Government Prioritizes Problems.* Chicago: University of Chicago Press, 2005. (with Bryan D. Jones)
- *Policy Dynamics*. Chicago: University of Chicago Press, 2002. (Edited, with Bryan D. Jones)

SBR 0111224, July 1, 2001 to June 30, 2004. "Lobbying and Issue-Definition."
SBR 9905195, August 1, 1999 to December 31, 2000. "Collaborative Research on Lobbying."
These two awards have supported the lobbying and advocacy project. Results include:

- http://lobby.la.psu.edu
- *Lobbying and Policy Change: Who Wins, Who Loses, and Why*. Chicago: University of Chicago Press, 2009 (with Jeffrey M. Berry, Marie Hojnacki, Beth L. Leech, and David C. Kimball).
- Several works in progress

References

Baumgartner, Frank R., Suzanna L. De Boef and Amber E. Boydstun. 2008. *The Decline of the Death Penalty and the Discovery of Innocence.* New York, NY: Cambridge University Press.

Baumgartner, Frank R. and Bryan D. Jones. 2009. *Agendas and Instability in American Politics*. 2nd Edition. Chicago, IL: University of Chicago Press.

Bennett, W. Lance, Regina G. Lawrence, and Steven Livingston. 2007. *When the Press Fails: Political Power and the News Media from Iraq to Katrina*. Chicago, IL: University of Chicago Press.

Boydstun, Amber E. 2008. How Policy Issues Become Front-Page News: The Media Dynamics of Information Processing, Conflict Displacement, and Social Cascades. PhD diss., The Pennsylvania State University.

Flemming, Roy B., B. Dan Wood & John Bohte. 1999. Attention to Issues in a System of Separated Powers: The Macrodynamics of American Policy Agendas. *The Journal of Politics* 61: 76–108.

Gans, Herbert J. 2004. *Deciding What's News: A Study of CBS Evening News, NBC Nightly News, Newsweek, and Time*. Evanston, IL: Northwestern University Press.

Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2007. Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology & Politics* 4(4): 31-46.

Hopkins, Daniel and Gary King. Forthcoming. A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science*.

Iyengar, Shanto. 1991. *Is Anyone Responsible? How Television Frames Political Issues.* Chicago, IL: University of Chicago Press.

Jones, Bryan D. & Frank R. Baumgartner. 2005. *The Politics of Attention: How Government Prioritizes Problems.* Chicago, IL: The University of Chicago Press.

Jones, Bryan D., Tracy Sulkin and Heather Larsen. 2003. Policy Punctuations in American Political Institutions. *American Political Science Review* 97(1):151-169.

Jones, Bryan D., James L. True and Frank R. Baumgartner. 1997. Does Incrementalism Stem from Political Consensus or from Institutional Gridlock?" *American Journal of Political Science* 41(4):1319-1339.

Klebanov, Beata Beigman, Daniel Diermeier, and Eyal Beigman. 2008. Lexical Cohesion Analysis of Political Speech. *Political Analysis* 16 (4): 447-463.

Landauer, Thomas K. & Susan T. Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. *Psychological Review* 104:211-240.

Laver, Michael, Kenneth Benoit & John Garry. 2003. Estimating the Policy Positions of Political Actors Using Words as Data. *American Political Science Review* 97:311-331.

Lowe, Will. 2004. Content Analysis and Its Place in the (Methodological) Scheme of Things. *Qualitative Methods* 2 (1): 25-27.

Lowe, Will. 2008. Understanding Wordscores. *Political Analysis* 16 (4): 356-371.

Monroe, Burt L. and Ko Maeda. 2004. Talk's Cheap: Text-Based Ideal-Point Estimation. Paper presented at the Annual Summer Meeting for Political Methodology, Stanford University, July 29-31, 2004.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis* 16 (4): 372-403.

Pear, Robert. New Medicare Plan For Drug Benefits Prohibits Insurance. *New York Times*, December 7, 2003.

Simon, Adam F. & Michael Xenos. 2004. Dimensional Reduction of Word-Frequency Data as a Substitute for Intersubjective Content Analysis. *Political Analysis* 12:63-75.

Soroka, Stuart. 2002. *Agenda-Setting Dynamics in Canada.* Vancouver: University of British Columbia Press.

Wolfe, Michelle, Amber E. Boydstun and Frank R. Baumgartner. 2009. Comparing the Topics of Front-Page and Full-Paper Stories in the *New York Times*. Paper presented at the annual national conference for the Midwest Political Science Association, April 2-5, Chicago.

Wood, B. Dan, and Jeffrey S. Peake. 1998. The Dynamics of Foreign Policy Agenda Setting. *The American Political Science Review* 92 (1):173-84.

Woolley, John T. 2000. Using Media-Based Data in Studies of Politics. *American Journal of Political Science* 44 (1): 156–73.

AMBER E. BOYDSTUN

Assistant Professor of Political Science
University of California, Davis • Davis, CA 95618
Phone 530 752 0966 • Fax 530 752 8666 • aboydstun@ucdavis.edu

Professional Preparation

B.A., 1999, St. John's College. Mathematics and Philosophy.
M.A., 2004, The Pennsylvania State University. Political Science.
Ph.D., 2008, The Pennsylvania State University. Political Science.

Full Time Academic Appointments

2008– Assistant Professor of Political Science, University of California, Davis.

Publications

Baumgartner, Frank R., Suzanna Linn, and Amber E. Boydstun. The Decline of the Death Penalty: How Media Framing Changed Capital Punishment in America. 2009. In Schaffner, Brian F. and Patrick Sellers (eds), *Winning with Words: The Origins and Impact of Framing*. New York: Routledge.

Baumgartner, Frank R., Suzanna L. De Boef, and Amber E. Boydstun. *The Decline of the Death Penalty and the Discovery of Innocence*. 2008. New York: Cambridge University Press.

Dardis, Frank E., Frank R. Baumgartner, Amber E. Boydstun, Suzanna De Boef, and Fuyuan Shen. 2008. Media Framing of Capital Punishment and Its Impact on Individuals' Cognitive Responses. *Mass Communication and Society* 11 (2): 115-140.

Synergistic Activities

Comparative Policy Agendas Project Media Data Collection Session, June 17, 2009, Den Hague, The Netherlands. Meeting brought together a group of scholars from different countries who are working in tandem to collect media data using the common Policy Agendas Project Topic Codebook. Session involved refining strategies for implementing the coding scheme in a consistent manner across countries.

Senior Honors Thesis Advisor for Laura Britton, 2008–2009, UC Davis. Laura's thesis, entitled "Media Framing of Female Candidates for President and Prime Minister: An International Comparison, and its Implications," involved collecting and content analyzing hundreds of newspaper stories covering presidential and prime minister elections in France, Ireland, New Zealand, and the US.

Doctoral Dissertation Improvement Grant, National Science Foundation Political Science Program, "Agenda-Setting and Issue-Framing Dynamics in Front-Page News." Grant No. SES-0617492, $10,907, July 1, 2006 to June 30, 2007 (Frank Baumgartner PI). Project produced an automated downloading software program that assisted in the retrieval of more than 30,000 news stories. Data collection involved training and overseeing fifteen undergraduate students in the process of using this automated download software and coding the stories retrieved using the Policy Agendas Project Topic Codebook.

TextTools Automated Classification Training Workshop, June 16, 2008, Penn State University.

Paul Wolfgang and colleagues from Temple University trained participants in the use of TextTools and in the best methods for iterated bootstrapping to maximize the accuracy and efficiency of the computer-assisted classification process.

Automated Text Identification and Classification for the Social Sciences Workshop, August 15-17, 2007, Penn State University. Invitation-only workshop brought leading computer scientists together with political scientists and sociologists with extensive experience in creating and coding large-scale text databases to learn about the latest computer science research in the area of automated and computer-assisted content analysis.

Collaborators

Frank Baumgartner (UNC Chapel Hill), Frank Dardis (Penn State), Rebecca Glazier (University of Arkansas, Little Rock), Suzanna Linn (Penn State), Matthew Pietryka (UC Davis), Betsy Sinclair (University of Chicago), Holloway Sparks (Emory University)

Graduate Advisors (PhD Committee at Penn State University)

Frank Baumgartner, UNC Chapel Hill, Chair (Formerly, Penn State University)
Suzanna Linn, Penn State University
Marie Hojnacki, Penn State University
Eric Plutzer, Penn State University
John McCarthy, Penn State University

Thesis Advisees

None

(Revised August 2009)

FRANK R. BAUMGARTNER
Richard J. Richardson Distinguished Professor of Political Science
University of North Carolina, Chapel Hill • Chapel Hill, NC 27599-3265
Phone 919 962 0414 • Fax 919 962 0432 • frankb@unc.edu

Professional Preparation (Education)

B.A., 1980, The University of Michigan. Political Science and French. Phi Beta Kappa
M.A., 1983, The University of Michigan. Political Science
Ph.D., 1986, The University of Michigan. Political Science

Full Time Academic Appointments

2009– Richardson Distinguished Professor of Political Science, UNC Chapel Hill
1998–09 Penn State University (Professor 1998–07; Miller-LaVigne Professor 2007–09)
1998–99 California Institute of Technology, Visiting Professor
1987–98 Texas A&M University (Assistant 1987–92; Associate 1992–97; Professor 1997–98)
1986–87 The University of Iowa, Visiting Assistant Professor

Selected Publications Related to this Grant Proposal

The Decline of the Death Penalty and the Discovery of Innocence. New York: Cambridge University Press, 2008 (with Suzanna L. De Boef and Amber E. Boydstun). (Kammerer Award, best book in US public policy, American Political Science Association, 2008)

Media Framing of Capital Punishment and Its Impact on Individuals' Cognitive Responses. *Mass Communication and Society* 11, 2 (2008): 115–140. (with Suzanna De Boef, Amber E. Boydstun, Frank E. Dardis, and Fuyuan Shen)

The Politics of Attention: How Government Prioritizes Problems. Chicago: University of Chicago Press, 2005. (with Bryan D. Jones)

Policy Dynamics. Chicago: University of Chicago Press, 2002. (Edited, with Bryan D. Jones)

Agendas and Instability in American Politics. Chicago: University of Chicago Press, 1993. (with Bryan D. Jones) (Wildavsky Award, book of lasting impact, 2001); 2nd edition 2009.

Other Significant Publications

A General Empirical Law for Public Budgets: A Comparative Analysis. *American Journal of Political Science*, forthcoming, October 2009. (Bryan D. Jones, Frank R. Baumgartner, Christian Breunig, Christopher Wlezien, Stuart Soroka, Martial Foucault, Abel François, Christoffer Green-Pedersen, Peter John, Chris Koske, Peter B. Mortensen, Frédéric Varone, and Stefaan Walgrave)

Punctuated Equilibrium in Comparative Perspective. *American Journal of Political Science*, 53, 3, (July 2009): 602–619. (Frank R. Baumgartner, Christian Breunig, Christoffer Green-Pedersen, Bryan D. Jones, Peter B. Mortensen, Michiel Neytemans, and Stefaan Walgrave)

Basic Interests: The Importance of Groups in Politics and in Political Science. Princeton: Princeton University Press, 1998. (with Beth L. Leech)

Synergistic Activities

"Database Development for the Study of Public Policy," NSF grant # SBR–0111611 for $690,719 covering the period from January 1, 2002 to December 31, 2007, with Bryan D.

Jones. Information concerning our project, as well as all of the data we have collected, is available at the Policy Agendas web site: www.policyagendas.org.

National Science Foundation, "New Computer Science Applications in Automated Text Identification and Classification for the Social Sciences." Grant # SES 0719703, $55,722, September 1, 2007 to August 31, 2008. PI, with John McCarthy.

National Science Foundation, "Lobbying and Issue-Definition." Grant # SBR 0111224, $235,930, July 1, 2001 to June 30, 2004. Principal Investigator. Co-Investigators are: Jeff Berry, Marie Hojnacki, Beth Leech, and David Kimball.

National Science Foundation, "Collaborative Research on Lobbying." Grant # SBR 9905195, $80,569, August 1, 1999 to December 31, 2000. Principal Investigator. Co-Investigators are: Jeff Berry, Marie Hojnacki, Beth Leech, and David Kimball.

Collaborators within the Past 48 Months

John McCarthy, Marie Hojnacki, Suzanna De Boef, Frank Dardis, Fuyuan Shen (Penn State), Bryan Jones (Washington), Amber Boydstun (UC Davis), Jeffrey Berry (Tufts), David Kimball (Missouri) Tim LaPira (College of Charleston), Beth Leech (Rutgers), Christine Mahoney (Syracuse), James True (Lamar) John Wilkerson (Washington) David Lowery (Leiden), Virginia Gray (North Carolina), Jim Stimson (North Carolina), Christian Breunig (Washington), Martial Foucault (Montreal), Abel François (Strasbourg), Christoffer Green-Pedersen (Aarhus), Peter John (Manchester), Chris Koske (Washington), Peter B. Mortensen (Aarhus), Stuart Soroka (McGill), Frédéric Varone (Geneva), Stefaan Walgrave (Antwerp), Michiel Neytemans (Antwerp), Chris Wlezien (Temple), Joe McGlaughlin (Temple), Andrew W. Martin (Ohio State), Heather Larsen-Price (Memphis), Trey Thomas (Texas), Ed Walker (Vermont)

Thesis Advisees (PhD committees chaired since 1997)

Amber Boydstun (Ph.D., Penn State; 2008; currently at UC Davis)
Christine Mahoney (Ph.D., Penn State; 2006; currently at Syracuse)
Beth Leech (Ph.D., Texas A&M, 1998; currently at Rutgers)
Michael MacLeod (Ph.D., Texas A&M, 1998; currently at Forestar Research)
Doris McGonagle (Ph.D., Texas A&M, 1998; currently at Blinn College)
James True (Ph.D., Texas A&M, 1997; currently at Lamar University)
Total graduate advisees since 1997: 6

Graduate Advisors (PhD Committee at the University of Michigan)

Roy Pierce, University of Michigan, Chair (deceased)
Jack L. Walker, Jr., University of Michigan (deceased)
Joel D. Aberbach, UCLA (Formerly, University of Michigan)

(Revised August 2009)

PAUL A. T. WOLFGANG

PROFESSIONAL PREPARATION

INSTITUTION	MAJOR/AREA OF STUDY	DEGREE	YEAR
University of Pennsylvania	Electrical Engineering	BS	1967
University of Pennsylvania	Computer and Information Science	30 credit hours of graduate study	1967-1968

ACADEMIC & PROFESSIONAL APPOINTMENTS

1999-Present	Instructor, Computer and Information Science, Temple University
1985-1999	Adjunct Instructor, Computer and Information Science, Temple University
1998-2001	Embedded Software Engineer, The Boeing Company
1986-1998	Manager, Software Engineering, The Boeing Company
1977-1986	Lead Scientist, Computer Sciences Corporation
1972-1977	Senior Associate, Analytics Inc.
1968-1972	Research Scientist, Pennsylvania Research Associates, Inc.

PUBLICATIONS

Koffman and Wolfgang. *Objects, Abstraction, Data Structures and Design using Java*. John Wiley and Sons, Inc. 2005. ISBN 0-471-46756-1

Koffman and Wolfgang. *Objects, Abstraction, Data Structures and Design using Java Version 5.0*. John Wiley and Sons, Inc. 2005. ISBN 0-471-69264-6.

Koffman and Wolfgang. *Objects, Abstraction, Data Structures and Design using C++*. John Wiley and Sons, Inc. 2006. ISBN 0-471-46755-3.

Wolfgang & Song "Integration of the STL and the Microsoft Foundation Class" *SIGPLAN Notices* vol. 34, no. 6 (June 1999).

Friedman & Wolfgang "Choosing Ada Tasking Models for Real-Time Systems" *Defense Electronics* vol. 19, no. 4 (April 1987)

SYNERGISTIC ACTIVITIES

Pennsylvania Policy Database

The goal of the Pennsylvania Policy Database is to establish the first comprehensive state database that will allow researchers, students, teachers, policy makers, and the general public to track and analyze policy in Pennsylvania through bills, statutes, appropriations, hearings, newspapers, opinion polls, executive orders, and court decisions. This database is modeled on the national Policy Agendas project hosted at the University of Washington. This project is funded by the Pennsylvania State Legislature, and is a cooperative effort of several Pennsylvania universities, with Temple in the lead. My contribution has been to adapt the software originally written for the national website to the Pennsylvania data. I have developed software to apply automatic classification algorithms to the data to reduce the manual labor involved, but preserving the essential classification properties. This software is also being used at Pennsylvania State University and the University of Washington.

COLLABORATORS

Jason Bossie, Carnegie Mellon University
Justin Gollob, Mesa State University
Jay Jennings, Temple University
Elliot B. Koffman, Temple University
Joseph P. McLaughlin, Jr., Temple University
John Wesley Leckrone, Weidner University

Michelle Wolfe

University of Texas at Austin · Department of Government
1 University Station, A1800 · Austin, TX 78712-0019
wolfemi@mail.utexas.edu

Professional Preparation (Education)

B.A., 2003, University of Washington, Political Science
2005-2008, Doctoral Student, University of Washington, Political Science
2008-Present, Doctoral Student, University of Texas at Austin, Government

Selected Publications Related to this Grant Proposal

"Public Policy and the Mass Media: An Information Processing Approach" with B.D. Jones. Forthcoming. In K. Voltmer and S. Koch-Baumgarten (eds.), *Public Policy and the Mass Media: Influences and Interactions*. London, UK: Routledge.

"The Institutionalization of Environmental Attention in the United States and Denmark: Multiple versus Single-Venue Systems" with Christoffer Green-Pedersen. Forthcoming. *Governance*.

Other Significant Publications

"Is Urban Politics a Black Hole? Analyzing the Boundary Between Political Science and Urban Politics" with J. Sapotichne and B.D. Jones. 2007. *Urban Affairs Review*, 43(1): 76-106.

Synergistic Activities

2005-Present Research Assistant to Professor Bryan D. Jones, University of Washington and University of Texas at Austin

- Project Manager for the Policy Agendas Project
- Supervise data collection and content coding of multiple datasets

2007-2008 Teaching Assistant for Undergraduate Fellows Program, Center for American Politics and Public Policy, University of Washington

- Assist in the conceptualization, design, data collection, and analysis of Undergraduate Fellows' research projects
- Supervise policy content coding projects

Collaborators within the Past 48 Months

Frank R. Baumgartner, Bryan D. Jones, Christoffer Green-Pedersen, Samuel Workman, Amber E. Boydstun, Ashley E. Jochim

Thesis Advisees

None

Graduate Advisors (University of Texas at Austin)

Bryan D. Jones, J. J. "Jake" Pickle Regents Chair in Congressional Studies

Maxwell E. McCombs, Jesse H. Jones Centennial Chair in Communication and Professor of Government

JONATHAN W. MOODY
Ph.D. Candidate
Pennsylvania State University, University Park • University Park, PA 16802-5200
Phone 901 485 8883 • Fax 814 863 8979 • frankb@unc.edu

Professional Preparation (Education)

B.A., 2005, The University of Mississippi. Political Science and Journalism. Phi Beta Kappa.
M.A., 2008, The Pennsylvania State University. Political Science
Ph.D., 2011 (pending), The Pennsylvania State University. Political Science

Full Time Academic Appointments

None

Selected Publications Related to this Grant Proposal

None

Other Significant Publications

Linn, Suzanna, Jonathan Moody and Stephanie Asper. 2009. "Explaining the Horse-Race of 2008." *PS: Political Science & Politics*. 42(3): 459-465.

Synergistic Activities

Research Assistantship: Pennsylvania Policy Database Project (supervisor Frank Baumgartner) – Supervised the collection and construction of the Governor's News Clippings dataset. Tasks included: newspaper article abstracting, topic coding, data cleaning, and database management.

Technical Skills: TextTools – Development of strategies for the application of the TextTools Automated Classification Software. Applications to: Pennsylvania Policy Database Project's Governor's News Clippings dataset; 1996-2006 *New York Times* dataset (preliminary).

Collaborators within the Past 48 Months

Suzanna Linn

Thesis Advisees (PhD committees chaired since xx date)

None

Graduate Advisors (PhD Committee at the Pennsylvania State University)

Suzanna Linn, Pennsylvania State University, Chair
Frank R. Baumgartner, University of North Carolina-Chapel Hill, special co-chair
Christopher Zorn, Pennsylvania State University
Eric Plutzer, Pennsylvania State University

(Revised August 2009)