The Automated Coding of Policy Agendas: A Dictionary-Based Approach*

Quinn Albaugh[1], Julie Sevenans[2], Stuart Soroka[1] and Peter John Loewen[3]

[1]Department of Political Science, McGill University
[2]Department of Political Science, University of Antwerp
[3]Department of Political Science, University of Toronto

Abstract. Until recently, the Policy Agendas community has focused on human coding texts using the policy agendas codebook. When researchers have needed to code more texts than is feasible or cost-effective, they have turned to machine learning methods. In contrast, we propose an automated dictionary-based content analysis approach for replicating the policy agendas codes for English and Dutch texts. We validate the dictionaries several ways. For the English-language dictionary, we use data from Prime Minister's Question Time (from the United Kingdom) dataset and the State of the Union speeches (from the United States), both human-coded for each country's respective codebooks. For the Dutch-language dictionary, we compare the results to media content and party manifestos from Belgium previously coded using the Policy Agendas codes. Results suggest that these two dictionaries may produce valid, reliable and comparable measures of policy agendas, and that may be is possible to create similar dictionaries for other languages.

*Prepared for delivery at the 6th annual Comparative Agendas Project (CAP) conference, Antwerp, June 27-29, 2013. Please direct correspondence about this paper to Quinn Albaugh at `quinn.albaugh@mail.mcgill.ca` .

(Re-)Assessing Automated Content Analysis for Policy Agendas

The Policy Agendas community has invested countless hours in topic-coding legislation, legislative debates, judicial decisions, media content, and public opinion. This effort has quite clearly been productive. There now is a vast and valuable literature not just on punctuated equilibria in policymaking, but on a wide array of links between the agendas of publics, media and policymakers. We now know more about when parties react to publics, about when media drive policymaking, about how framing and policy communities shift not just the nature of policy debate, but the content of policy itself – and about the institutional environments in which each of these is more or less likely.[1]

The data-gathering effort that was initiated by the cross-national search for punctuated equilibria in politics has, in short, produced a considerable body of work on a range of important topics in political science and public policy. That said, we have only just begun to take advantage of what is an increasingly accessible, and massive, body of textual data on policy processes around the world. Even relatively simple codes, such as what are known in the Policy Agendas community as "Major Topic" codes, are time-consuming to generate. Human coding also raises issues of intercoder reliability – a major issue in some countries' efforts to capture policy agendas.[2]

One response amongst those interested in capturing attentiveness to topics in legislative (textual) data has been to focus on automated approaches. There is in fact a growing body of work on various methodologies for automated content analysis. (We will not discuss these in detail here, but note that there are several useful reviews, including, e.g., Grimmer and Stewart (2013).) There is in addition a particularly valuable body of work on the potential for "supervised learning" approaches aimed specifically at the Policy Agendas project (Purpura and Hillard 2006; Hillard et al. 2007, 2008).[3] This approach quite clearly has potential where topic classification is concerned.[4] Here, we want to take a step back, however – we want to re-assess the potential for very simple, dictionary-based approaches to topic coding.

Given that many different and sophisticated supervised learning approaches exist for automated coding, why reconsider what seems to be a rather blunt, dictionary-based approach? We see several potential advantages. The most important is that dictionary based approaches are completely and totally clear about what they are counting. There is no algorithm working with an unknown set of words. This may lead to worse outcomes coding-wise, or not, we are as yet unsure. But there is no mystery in the dictionary-based approach – you get a frequency count

[1] The literatures are vast, but see Baumgartner and Jones (2005); Baumgartner et al. (2009, 2011); Jones et al. (2003, 2009); Jones and Baumgartner (2012)

[2] Minor topic codes are even more difficult. In the Canadian data, for instance, coders worked for up to three months before achieving 95% intercoder reliability in minor topic codes.

[3] Also see the information on software for text classification on the Comparative Agendas Project website, http://www.comparativeagendas.info.

[4] And there is a related and increasingly sophisticated literature on automated coding generally, as well as for project closely allied with Policy Agendas work. See (Bond et al. 2003; Budge and Pennings 2007; Collingwood and Wilkerson 2012; Conway 2006; Diermeier et al. 2012; Farnsworth et al. 2010; Hart 1984, 2005; Hopkins and King 2010; Klebanov et al. 2008; König et al. 2010; Lowe 2008; Proksch and Slapin 2009, 2011, 2012; Quinn et al. 2010; Schrodt et al. 1994; Slapin and Proksch 2008, 2010; Soroka 2006, 2012; Vliegenthart and Walgrave 2011; Young and Soroka 2012; Yu et al. 2008).

based on a list of words, and that's all.

Dictionary-based systems may also be relatively effective where the coding of legislative data are concerned, because policy topics are actually relatively easily identified using a finite set of keywords. This too may be a subject for further discussion. But our own preliminary work suggests that economic policy does tend to include some very standard economic keywords; the same is true for environmental policy; and so on. The end result is that sophisticated approaches, which are necessarily somewhat opaque in terms of the coding process, may not produce systematically better codes than simple dictionary counts.

There is one major caveat: dictionary-based approaches will be relatively good at identifying the various topics mentioned in a given text, but they may not be especially reliable at assigning just one topic per text. This is a serious limitation, given that Policy Agendas work has tried to assign single topics whenever possible. But the single-topic approach may not be the best way to capture shifts in policy attentiveness; it certainly is not the best way to capture policy framing, since allowing for multiple topics allows us to capture the various ways in which attentiveness to one domain may be linked to attentiveness to others. We do not wish to justify a dictionary-based approach simply by abandoning single topics, mind you; indeed, we look below at the potential for a simple dictionary system to assign single topics. But we do wish to note that where single topics are a lower priority, dictionary coding may be particularly productive.

We also want to demonstrate a fact that is widely acknowledged but often ignored in automated content analysis: it usually works better when there is more data (i.e., words). Capturing topics in a 100-word corpus will in all likelihood be more reliable than capturing words in a 20-word corpus.[5] More importantly, however, capturing general trends in a 50,000-article sample can be highly reliable, even as the coding of each individual item is noisy. This clearly is not the primary goal for assigning topics to Policy Agendas data. In most cases, researchers want each individual item to be properly coded. But there are instances in which general trends are important, and in these instances dictionary-based – and indeed all automated coding – might be especially useful.

Of course, all of this is true only if dictionary-based systems work. We examine that possibility below. First, we discuss the construction of our topic dictionaries. We then introduce the various content analytic datasets we use to test the performance of these dictionaries against human coding. Analyses make clear that there is room for improvement, but that there is potential as well, particularly for certain topics, and particularly when we are interested in broad trends rather than item-by-item codes.

Dictionary Construction

So far, we have two dictionaries, one in English and one in Dutch, that cover the major topic codes under the Policy Agendas codebooks. We focus on major topic codes only, not minor topics, since major topic codes are sufficient for most analyses. But major topics are also

[5] Though this may vary - legislative titles may actually be easier to code when they include fewer words, that is, when the title is more direct.

quite clearly easier to capture, no matter the method of automation; and when necessary many useful minor codes, like inflation and unemployment, can be drawn from the bag of words used to represent their corresponding major topics.

Both dictionaries are configured in .lcd format for use with Lexicoder, a very simple Java-based program for automated text analysis developed by Lori Young and Stuart Soroka and programmed by Mark Daku. For those who have access to other automated content analysis software, however, it is not be difficult to convert the .lcd dictionaries to a different format (using any plain-text editing software).

The English-language dictionary came first. It was built in several stages over a number of years. Originally, it appeared as the Lexicoder Topic Dictionary, which has itself gone through a number of iterations.[6] Last fall, it was revised to more directly correspond with the Policy Agendas codes by going through the Canadian Policy Agendas codebook and copying the keywords into the dictionary, while also adding synonyms and related terms wherever possible. During this phase of the dictionary construction, both American and Canadian terms ("7th Grade" vs. "Grade 7") and American and Canadian spellings ("labor" vs. "labour") were included.[7] The entries in the English dictionary are not necessarily whole words or phrases but are often stemmed. For example, instead of including the words economy, economics and so forth, the dictionary includes the stem ECONOM-. To avoid picking up patterns of characters that might occur as parts of other words than the target, the English dictionary generally has spaces before each entry.

The terms from the English dictionary were then translated into Dutch as a starting point for the Dutch dictionary. During this process, the translation was adapted to better correspond with the Belgian Policy Agendas codebook, and some additional terms were added to both dictionaries based on a consultation of the Belgian codes. Notably, the Dutch-language dictionary does not include nearly as many spaces at the beginning or the end of words, since Dutch (like a number of other closely related Germanic languages) frequently builds new words by compounding several roots together. Since the corpora the Dutch-language dictionary was originally tested on were missing accented characters, it was revised to be more robust to situations in which the accented characters are missing (or, as has been the case with some Belgian data, replaced with a "?"), largely by stemming.

Additional terms were added to the English-language dictionary based on items that appeared in the Dutch-language dictionary and based on a similar close reading of the U.S. and U.K. Policy Agendas codebooks. Relatively few additional terms were taken from the other countries' codebooks, and great care was taken to ensure that adding terms did not diverge consistently from the other countries' codebooks. In order to do this, some topics were separated out for the purposes of the dictionary. In the current version of the dictionary, immigration is a separate category, and it must be merged with Topic 2 (Civil Rights) in Canada and the U.K. or Topic 5 (Labour and Employment) in the U.S. Similarly, fisheries is a separate category in the dictionary because the Canadian codebook treats it as separate from agriculture, unlike the

[6] A similar topic dictionary based on Nexis keywords was used in Farnsworth et al. (2010).

[7] This initial version of the dictionary was tested against the subject headings written in the Oral Questions section of Hansard for the Canadian House of Commons as a preliminary test of its validity. Results are available on request. However, this test has not been applied to the most recent version of the dictionary.

other two countries.

Since the Belgian Policy Agendas codebook differs considerably from the American, Canadian and British codebooks,[8] we decided that the best course of action was to develop dictionaries that matched up with the already existing codebooks for each country, rather than trying to standardize the codes across countries ourselves.

(Both topic dictionaries are freely available at lexicoder.com. Future versions will be available there as well; we discuss our aim to "crowd-source" the development of multilingual dictionaries in the concluding section.)

Corpus Selection

To test these dictionaries, we draw upon a wide variety of different corpora, including legislative debates, television and newspaper content, party manifestos and major speeches. We did not selected texts here based on theoretical interests. Instead, we chose texts based on two criteria: (1) whether they were in one of the two languages for which we have automated dictionaries and (2) whether they had been (in all but one case) human-coded using one of the Policy Agendas codebooks. The human-coded texts are necessary because our main goal here is to compare our automated coding results with human coding – this will be our test of validity for the dictionaries.

To test the Dutch-language dictionary, we rely on two sets of data sources taken from the Belgian Policy Agendas team. The first is media content from 2004-2008, including newspaper articles from *De Standaard* (N = 12,584) and television transcripts (N = 91,170) from Vlaamse Televisie Maatschappij (VTM), the major commercial television broadcaster in Flanders, and Vlaamse Radio- en Televisieomroeporganisatie (VRT) TV-1, the national public service broadcaster for the Flemish Region and Community in Belgium. In analyzing this set of data, we use the dictionary to code individual news stories.

For the English-language dictionary, we draw on from the United States and the United Kingdom, both of which have teams that have posted a wide variety of datasets coded for Policy Agendas; as a result, we were able to find corpora suitable for setting up similar tests to the ones we did with the Dutch-language dictionary. To parallel the Flemish media content, we draw upon the Prime Minister's Question Time data from 1997-2008 posted on the U.K. Policy Agendas project website (N = 9,062).[9]

In addition, since a substantial number of Policy Agendas datasets are broken down by quasi-sentences to capture the relative weight of topics within a long text, we designed a different test to assess how well the dictionary would perform. Since quasi-sentences are quite short, and since past experience with dictionary-based approaches suggests that short texts produce more noise than long texts, we instead compare the human-coding approach, aggregating codes

[8] For example, the Belgian codebook has a separate topic for Immigration that does not exist for these other countries (though would probably be useful), and it classifies a number of issues in other categories; for example, abortion is a health care issue under the Belgian codebook but a civil rights and liberties issue under the Anglosphere codebooks.

[9] The dataset is available online at `http://policyagendasuk.wordpress.com/datasets/`.

at the quasi-sentence level to produce relative proportions for the entire document, with the dictionary-based approach, at the level of the entire document. For this, we took State of the Union speeches from the United States from 1946-2012 (N = 66).[10]

Note that the U.S. and U.K. datasets provide relatively strong tests of how well the dictionary captures the major topics in a cross-national sense because the dictionary was originally designed to be used on data from a different country entirely (Canada).

Analysis

Recall that we are interested in the performance of dictionaries at two different levels of aggregation. On the one hand, we are interested in the reliability of the dictionaries at identifying topics in single items. On the other hand, we are interested in the reliability of dictionaries at capturing topics across a range of items. The first corresponds to many Policy Agendas researchers' interests in individual titles of legislative, sections of legislative debates, court decisions, and so on. The second comes closer to what we see in manifesto-level data from the Comparative Manifestos Project; or over-time trends based on aggregated Policy Agendas data. It is also closer to the way in which most Policy Agendas researchers deal with media content, aggregated weekly or monthly. (Though note that whereas aggregate-level CMP or Policy Agendas data are typically aggregations of item-level codes, we are simply going to count words across all items simultaneously.)

Item-Level Codes

We begin with assigning codes to individual items. For the English dictionary we do this using publicly available data on Prime Minister's Questions from the United Kingdom. We kept only those cases that already had coding for one of the major topics. Any cases with the code "99" (Miscellaneous/Other) are not included in the analysis. In the case of Prime Minister's Questions, this was not an issue, as all the questions were coded with one of the 21 major topics. (For the sake of consistency in subsequent analyses, we also drop cases coded for any topics other than the 21 analyzed here.)

[Table 1 about here.]

In order to maintain comparability with most Policy Agendas datasets, we had to devise a way to assign one and only code to each case using the automated coding scheme. To do this, we first selected the topic most frequently mentioned within each text, which here includes both the question and the answer text (where an answer was available).[11] Table 1 displays the number of Policy Agendas topics identified in each question-answer combination. Zero indicates

[10] The dataset is available online at `http://www.policyagendas.org/page/datasets-codebooks`.

[11] Note that we ignore the fact that some topics have more keywords than others. In the English-language dictionary, the number of keywords range from 23 to 128. It may be that these topics are more like to produce plurality-winning totals. It may also be that some topics simply need fewer words to be identified. So there is only one word for fishing, but anything about fishing includes that word. It follows that the number of items

a null result for the automated coding - this means none of the dictionary entries associated with any of the Policy Agendas topics were mentioned at all anywhere in the text. A result of one indicates a clear winner among the topics. That is, one topic is clearly mentioned more frequently than the others. The remaining numbers indicate ties–a result of two indicates a two-way tie, three indicates a three-way tie, and so forth.

Any results other than one are a problem, either because we cannot attribute a code at all (9.8% of all items), or because there are ties between multiple topics (15% of all items). Null results indicate much bigger problems with the dictionary than do ties – it is relatively easy to find a way to resolve ties, after are. But the dictionary is relatively good at coming up with at least one topic for each case; whether these topics correspond well to the human-coded Policy Agendas topics is another matter.

Let us consider first the uncoded cases. As a diagnostic, we checked to see which kinds of questions – based on human-coded topics – were not coded with the dictionary. The clear majority (58 percent) of question-answer pairings not coded at all with the automated coding fell under Major Topic 20: Government Operations.[12] We take this as a pointer for future work: there are words missing from the dictionary's Government Operations category. That said, this is one topic where dictionary terms likely differ considerably from one country to another. For example, in the United Kingdom, this topic often covers the House of Lords, the monarchy and other institutions that might not be present (or present in the same ways, given Canada's different relationship with the monarchy) across other countries. So improving dictionary performance for this topic must proceed on a country-by-country basis.

Table 1 also indicates that the automated coding only produced a tie between cases for 16 percent or so of the question-answer pairings. This suggests that, usually, just taking the topic mentioned most frequently in the text may be enough, but it's not *always* sufficient. To break these ties, we took the topic first mentioned within each text. This is a fairly intuitive measure. Breaking ties using the first mentioned topic may also have a clear advantage over human coding, since human coders will consider multiple possible codes as valid for a single case, and different coders are likely to break ties in different ways. Taking the topic first mentioned in the text, as indicated with these dictionaries, should guarantee a higher standard of reliability for choosing between topics than relying on whatever human coders pick. (Though note that this line of reasoning raises another issue: perhaps human codes are not what we should be trying to reproduce.)

[Table 2 about here.]

Just how well does the dictionary-based coding hold up against the human coding? Table 2 presents a series of analyses that compare the automated coding against the human coding, using two sets of marginal probit regressions that use automated coding to predict human

in a dictionary category may matter. Preliminary evidence from the English-language dictionary suggests that topics with more keywords perform somewhat better overall; however, more study is needed.

[12] Full results available upon request. The category that came in second for most cases not coded with the dictionary was Major Topic 19: International Affairs and Foreign Aid, which accounted for only 9 percent of the question-answer pairings not coded with the dictionary.

coding. The first set of marginal probit regressions (listed in the column labelled with "Raw word counts") takes the dictionary's count of the number of times each topic was mentioned in the text and uses it to predict whether the text was human-coded for that topic. The second set (listed in the column labelled with "Mutually-exclusive topics") takes a series of dummy variable coded "1" for each if the most mentioned topics in that text using the dictionary was the one in question, breaking ties using the first-mentioned topic, and uses it to predict whether each text was human-coded for that topic. (From now on, we will simply refer to this as either our mutually-exclusive topics scheme, or as our approximation of the Policy Agendas coding scheme.)

For each set of marginal probit regressions, we first display the Pseudo-R^2, followed by the marginal probit coefficient (indicated by *dfdx*). (Note that we use Pseudo-R^2 as a simple, readily interpretable measure of model fit – a signal that the models well, or not so much. We readily acknowledge that this is an imperfect measure of the fit of probit regressions. For our purposes here, however, it should suffice.) Note that the marginal probit coefficients are not directly comparable between the first and second set of models – the independent variable in the first set takes on whole numbers that can run much higher than one, while, in the second set, the independent variable is coded as binary, either zero or one. As a result, the coefficients are much higher for the second set.

Table 2 make clear that the performance of the dictionary differs considerably by topic. The mean Pseudo-R^2 for the raw word counts is 0.30, which is promising but suggests plenty of room for improvement. However, some topics perform quite well. In particular, healthcare, agriculture and education are quite well-captured by the dictionary, with Pseudo-R^2 values that are close to 0.6. Note that, given the likelihood that there is some amount of error in human coding, it is not clear that we would want the dictionary coding to perfectly predict the human coding; however, codes that are measuring the same underlying phenomena should correspond fairly well to one another. These three topics may not need much revision. By contrast, we also have three topics that performed particularly poorly, with Pseudo-R^2 values that are practically zero: International Affairs and Foreign Aid, Government Operations and Public Lands and Water Management. As discussed above, the poor results for Government Operations are not very surprising, given that the dictionary used here attempts to cover the United States, Canada and the United Kingdom, whereas dictionaries adapted for each country would not be likely to share many of the same terms.

If you compare the first two columns of Table 2, you can see that, for many of the topics, more variance in the human coding is explained when you dummy out the coding based on the which topic is most frequently mentioned in the text, breaking ties by looking at the first-mentioned topic. In fact, the mean increase in the Pseudo-R^2 is about 0.03. This is unsurprising, given that that this dummy variable more closely mirrors the Policy Agendas codebook, which restricts us from assigning multiple topics. However, for four topics (Labour, Environment, Social Welfare and Banking, Finance and Domestic Commerce), the Pseudo-R^2 actually *decreased* when moving from the raw dictionary counts to our approximation of the (mutually-exclusive) Policy Agendas codes. Clearly, at times, the constraint of being forced to pick a single topic code can have its downsides.

We can run similar tests on Dutch-language media from Belgium. Tables 3 and 4 display the number of topics for newspaper articles and television stories. Note that the mean word counts differ considerably between newspapers (mean = 288 words) and television (mean = 22 words). Surprisingly, the very short television blurbs had far fewer cases not coded by the dictionaries (17.43) in comparison with the newspaper articles (24.6 percent uncoded). However, the Dutch-language dictionary left far more cases uncoded with media content than the English-dictionary did with Prime Minister's Questions. This does not seem particularly surprising, as parliamentary debates may be particularly predisposed to having policy-related words in them, as opposed to media content.

[Table 3 about here.]

[Table 4 about here.]

Here, though, there was no single topic that appears to be behind the bulk of the uncoded cases. For both newspaper and television stories, the two topics that explained the most of this group were Criminality and Legal Issues (14 percent for newspaper and 12 percent of television stories) and Functioning of Democracy and Government (10 percent for newspaper and 14 percent for television stories). These two topics likely need revision; however, it is possible that these types of corpora are simply more difficult to code.

[Table 5 about here.]

[Table 6 about here.]

Likewise, we can replicate the two sets of marginal probit regressions used with the English-language version of the dictionary. Tables 5 and 6 show the same , with the "Raw word counts" column showing results allowing for multiple topics using the raw dictionary results, and the "Mutually-exclusive topics" column showing the results when we force the coding into a single topic. Overall, the results for the Dutch-language and English-language dictionaries are fairly similar, with a wide variety of range of results across topics. The mean Pseudo-R^2 was slightly higher for both Belgian TV (0.35) and Belgian newspapers (0.34) than for the Prime Minister's Questions (0.30). Notably, the Dutch-language dictionary's worst topics did not perform as poorly as the English-language ones. The two topics that stick out the most at the high end of the scale are Immigration and Education.

As with the English-language dictionary, forcing one single topic led generally led to an improvement in predicting the human coding. For Belgian TV, the mean improvement in Pseudo-R^2 was 0.10, with only three topics performing worse than just using the raw word count: Civil Rights and Civil Liberties, Immigration and the media code Fires and Accidents. For those topics that decreased, the decrease was much smaller than with the Prime Minister's Questions dataset. For *De Standaard*, mean improvement in Pseudo-R^2 was 0.06. Seven topics performed more poorly when forcing the coding into mutually-exclusive topics: the same three as Belgian TV, plus Environment, Foreign Trade, International Affairs and Foreign Aid and Public Lands

and Water Management. (Bear in mind that the Dutch-langauge dictionary has more topics than the English-language one, since the Belgian codebook adds major topics and also a variety of media codes.)

It may be that forcing a single topic code simply works better the shorter the text. In our data, note that it worked best – though only marginally so – with the extremely short Belgian TV dataset. The longer the text, the more it may make sense for us to take an approach that takes into account the relative weight of different topic codes within the same text. For this, we turn to datasets previously coded at the quasi-sentence level.

Aggregate-Level Codes

One key advantage of breaking down longer texts, such as party manifestos, court decisions or speeches, into quasi-sentences it that each quasi-sentence can have a single mutually-exclusive code. In this respect, dividing based on quasi-sentences helps maintain the Policy Agendas principle of one case, one code. However, dividing a text into quasi-sentences can itself be costly in terms of time. It might also lose important information, paricularly for work that focuses on issue framing, or issue co-occurrences. In short, we do not feel that the best route is to *always* search out small, single-code items.

Indeed, we suggest that, if the goal is to assess the relative weight of codes within long texts, or across periods (or parties, or countries) with multilple texts apiece, it is neither necessary nor preferrable to divide text into quasi-sentences. Instead, we should focus on dictionary counts across combined items. Since longer texts (or combined items) are likely to have more words from the same documents, it is usually necessary to adjust the raw word count by taking the percentage of words from each topic within the document or the number of topic words per thousand words to compensate. But this is of course relatively easy to manage.

To explore this possibility, we draw here upon the U.S. State of the Union speeches from 1946-2012. We take the mean number of words from each topic, and make an adjustment based on document length. We compared this automated measure with human coding in two ways. First, we take the correlation between the automated coding and the proportion of quasi-sentences human-coded for each topic within the entire speech. Second, we run simple ordinary least squares (OLS) regressions for each topic, using the mean word counts to predict the relative proportion of quasi-sentences human-coded. We display the results for both the correlation and the OLS regressions in Table 7.

[Table 7 about here.]

Of course, the correlation is simply the square root of the R^2; however, we display both to provide alternative ways of looking at the same information. As with the item-level coding, sometimes the automated coding performs quite well, while in other cases the automated results have practically no explanatory value. Again, some topics, like Education, Energy, Healthcare and Agriculture perform consistently quite well for the English-language dictionary, while others, like Social Welfare, Foreign Trade, Government Operations and Public Lands and Water

Management do not appear to be very useful at present. Indeed, the automated coding of Government Operations is effectively worthless in explaining the variance in the human coding.

The average R^2, taking into account all the topics, is roughly 0.41; however, if we omit the four worst topics, it rises to 0.51. This suggests how much these topics pull down the performance of the dictionaries overall. Furthermore, it suggests that using the aggregate-level approach on longer documents makes the automated coding results stronger, though it is not really appropriate to directly compare a Pseudo-R^2 from marginal probit regressions with a true R^2 from OLS regressions.

In terms of the rank order of topics by R^2, it seems worth noting that Macroeconomics does somewhat better using an aggregate-level coding scheme in longer documents than it did with item-level codes of shorter documents. This could be the product of the corpus selection; however, it could also be because words related to Macroeconomics frequently appear with other topics. In the automated coding, we no longer have to worry about cues for multiple topics appearing within the same quasi-sentence, since we are no longer concerned with maintaining mutually-exclusive codes. This is perhaps the clearest strength of the dictionary approach.

These results are illustrative for the English-language dictionary; however, while the specific results by topic are likely to be somewhat different for the Dutch-language dictionary, the general principle that automated coding can produce results that closely mirror human coding should still hold. The trick is figuring out how to optimize the dictionaries for each topic.

Discussion & Future Steps

Overall, the current versions of these dictionaries provide just a starting point for future work in dictionary-based automated coding of Policy Agendas topics. For the most part, the automated coding corresponds fairly well with the human coding, though it is far from perfect. We are as yet unsure of how best to characterize the relationships we find here. Recall that we should not expect the automated coding and the human coding to correspond perfectly. Human coders make mistakes. But they also have insights that automation cannot. It is possible that politicians, journalists or other figures drafting political texts will choose unusual ways of referring to topics – that only humans will idenfity. (This is we suspect as true for supervised learning approaches as it is for dictionary-based approaches.)

Importantly, while the two dictionaries discussed here are almost certainly not perfect, they are extremely consistent. The words do not change across documents, so we can be extremely confident that we know what the dictionary is doing regardless of what corpus we select. We see this as an important advantage.

We also see real potential for the admittedly simple, but we believe powerful, dictionary-based approach. Even our early efforts have produced highly reliable codes across many of the Policy Agendas "Major Topic" codes. These can improve with time and testing. And note that reliability is markedly better when we abandon the need to assign only single topics. This will not work for everyone, we admit. But it clearly would help for researchers interested in over-time trends in media content, for instance, where assigning single topics to each article is

neither required nor, we believe, desirable.

We also see real potential for the development of multi-lingual versions of the topic dictionarites used here. Towards that end, we have made current versions of both dictionaries readily available at lexicoder.com. We welcome any and all suggested revisions to the English- and Dutch-language versions; and we epscially welcome efforts to translate the dictionaries into other languages. (The ability to identify topics, even in languages you and your coders do not speak, is clearly attractive. It does require that native-speakers develop and share dictionaries, however; and that those dictionaries are actively tested by researchers familiar with the language at hand.)

Our own next steps will depend to some extent on what happens with the Policy Agendas Master Codebook – a project currently underway. In a number of respects, the proposed revisions to the Master Codebook mirror what we have already done with the dictionary, for example, by creating a separate topic for immigration issues. However, it should be relatively easy to modify the dictionary later to match the Master Codebook. There is, in sum, no reason to delay the development of topic dictionaries.

References

Baumgartner, Frank R., Christian Breunig, Christoffer Green-Pedersen, Bryan D. Jones, Peter B. Mortensen, Michiel Neytemans, and Stefaan Walgrave. 2009. "Punctuated Equilibrium and Institutional Friction in Comparative Perspective." *American Journal of Political Science* 53.

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

Baumgartner, Frank R, Bryan D Jones, and John Wilkerson. 2011. "Comparative Studies of Policy Dynamics." *Comparative Political Studies* 44(8):947–972.

Bond, Doug, Joe Bond, Churl Oh, J Craig Jenkins, and Charles Lewis Taylor. 2003. "Integrated Data for Events Analysis (IDEA): An Event Typology for Automated Events Data Development." *Journal of Peace Research* 40(6):733–745.

Budge, Ian and Paul Pennings. 2007. "Do they work? Validating computerised word frequency estimates against policy series." *Electoral Studies* 26(1):121–129.

Collingwood, Loren and John Wilkerson. 2012. "Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods." *Journal of Information Technology & Politics* 9(3):298–318.

Conway, Mike. 2006. "The Subjective Precision of Computers: A Methodological Comparison with Human Coding in Content Analysis." *Journalism & Mass Communication Quarterly* 83(1):186–200.

Diermeier, Daniel, Jean-François Godbout, Bei Yu, and Stefan Kaufmann. 2012. "Language and Ideology in Congress." *British Journal of Political Science* 42(1):31–55.

Farnsworth, Stephen, Stuart Soroka, and Lori Young. 2010. "The International Two-Step Flow in Foreign News: Canadian and U.S. Television News Coverage of U.S. Affairs." *The International Journal of Press/Politics* 15(4):401–419.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* .

Hart, R.P. 1984. *Verbal Style and the Presidency: A Computer-Based Analysis*. Human communication research series, Academic Press.

Hart, R.P. 2005. *Political Keywords: Using Language That Uses Us*. Oxford University Press.

Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2008. "Computer-Assisted Topic Classification for Mixed-Methods Social Science Research." *Journal of Information Technology & Politics* 4(4):31–46.

Hillard, Dustin, Stephen Purpura, John Wilkerson, David Lazer, Michael Neblo, Kevin Esterling, Aleks Jakulin, Matthew Baum, Jamie Callan, and Micah Altman. 2007. "An active learning framework for classifying political text." In *Presented at the annual meetings of the Midwest Political Science Association*.

Hopkins, Daniel J and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54(1):229–247.

Jones, BD, FR Baumgartner, C Breunig, C Wlezien, S Soroka, M Foucault, A François, C Green-Pedersen, C Koski, and P John. 2009. "A general empirical law of public budgets: A com-

parative analysis." *American Journal of Political Science* 53(4):855–873.

Jones, BD, T Sulkin, and HA Larsen. 2003. "Policy Punctuations in American Political Institutions." *American Political Science Review* 97(01):151–169.

Jones, Bryan D and Frank R Baumgartner. 2012. "From There to Here: Punctuated Equilibrium to the General Punctuation Thesis to a Theory of Government Information Processing." *Policy Studies Journal* 40(1):1–20.

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

König, Thomas, Bernd Luig, Sven-Oliver Proksch, and Jonathan B. Slapin. 2010. "Measuring Policy Positions of Veto Players in Parliamentary Democracies." In *Reform Processes and Policy Change: Veto Players and Decision-Making in Modern Democracies*, Thomas König, Marc Debus, and George Tsebelis, eds., pages 69–95, New York: Springer.

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

Proksch, Sven-Oliver and Jonathan B Slapin. 2009. "How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany." *German Politics* 18(3):323–344.

Proksch, Sven-Oliver and Jonathan B Slapin. 2011. "Parliamentary questions and oversight in the European Union." *European Journal of Political Research* 50(1):53–79.

Proksch, Sven-Oliver and Jonathan B Slapin. 2012. "Institutional Foundations of Legislative Speech." *American Journal of Political Science* 56(3):520–537.

Purpura, Stephen and Dustin Hillard. 2006. "Automated classification of congressional legislation." In *Proceedings of the 2006 international conference on Digital government research*, pages 219–225, Digital Government Society of North America.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54(1):209–228.

Schrodt, Philip A, Shannon G Davis, and Judith L Weddle. 1994. "Political Science: KEDS—A Program for the Machine Coding of Event Data." *Social Science Computer Review* 12(4):561–587.

Slapin, Jonathan B and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722.

Slapin, Jonathan B and Sven-Oliver Proksch. 2010. "Look who's talking: Parliamentary debate in the European Union." *European Union Politics* 11(3):333–357.

Soroka, Stuart N. 2006. "Good News and Bad News: Asymmetric Responses to Economic Information." *Journal of Politics* 68(2):372–385.

Soroka, Stuart N. 2012. "The Gatekeeping Function: Distributions of Information in Media and the Real World." *The Journal of Politics* 74:514–528.

Vliegenthart, Rens and Stefaan Walgrave. 2011. "Content Matters: The Dynamics of Parliamentary Questioning in Belgium and Denmark ." *Comparative Political Studies* 44(8):1031–1059.

Young, Lori and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." *Political Communication* 29:205–231.

Yu, Bei, Stefan Kaufmann, and Daniel Diermeier. 2008. "Classifying Party Affiliation from Political Speech." *Journal of Information Technology & Politics* 5(1):33–48.

Table 1: Number of Topics Coded using Dictionary, Prime Minister's Question Time (UK)

	Freq.	Percent
0	889	9.81
1	6739	74.37
2	994	10.97
3	311	3.43
4	94	1.04
5	31	0.34
6	2	0.02
7	1	0.01
8	1	0.01
Total	9062	100

Table 2: Predicting Human Coding with Automated Coding using Marginal Probit, Prime Minister's Question Time (UK)

	Raw word counts		*Mutually-exclusive topics*	
	Pseudo-R2	**dfdx**	**Pseudo-R2**	**dfdx**
Macroeconomics	0.33	0.03	0.38	0.44
Civil Rights	0.29	0.03	0.29	0.51
Healthcare	0.57	0.04	0.69	0.83
Agriculture & Fisheries	0.56	0.01	0.60	0.75
Labour	0.24	0.02	0.16	0.20
Education	0.59	0.02	0.66	0.73
Environment	0.42	0.02	0.38	0.63
Energy	0.40	0.01	0.53	0.67
Transportation	0.38	0.03	0.40	0.61
Crime, Law & Family Issues	0.40	0.04	0.47	0.61
Social Welfare	0.33	0.04	0.25	0.66
Housing & Community Development	0.31	0.04	0.32	0.82
Banking, Finance & Domestic Commerce	0.13	0.02	0.12	0.34
Defence	0.27	0.12	0.34	0.72
Space, Science, Technology & Communications	0.27	0.01	0.34	0.40
Foreign Trade	0.21	0.00	0.25	0.12
International Affairs & Foreign Aid	0.06	0.06	0.10	0.40
Government Operations	0.01	0.10	0.02	0.56
Public Lands & Water Management	0.00	0.01	0.01	0.06

Table 3: Number of Topics Coded using Dictionary, De Standaard (Belgium)

	Freq.	Percent
0	3008	24.61
1	7610	62.25
2	1201	9.82
3	277	2.27
4	80	0.65
5	39	0.32
6	8	0.07
7	0	0.00
8	1	0.01
Total	12224	100

Table 4: Number of Topics Coded using Dictionary, VRT and VTM Television (Belgium)

	Freq.	Percent
0	13055	17.43
1	54232	72.42
2	6599	8.81
3	886	1.18
4	108	0.14
5	6	0.01
Total	74886	100

Table 5: Predicting Human Coding with Automated Coding using Marginal Probit, De Standaard (Belgium)

	Raw word counts		*Mutually-exclusive topics*	
	Pseudo-R2	**dfdx**	**Pseudo-R2**	**dfdx**
Macroeconomics	0.35	0.01	0.35	0.46
Civil Rights	0.22	0.02	0.20	0.52
Healthcare	0.37	0.01	0.46	0.60
Agriculture & Fisheries	0.41	0.01	0.56	0.69
Labour	0.44	0.01	0.48	0.57
Education	0.60	0.00	0.67	0.57
Environment	0.47	0.01	0.40	0.61
Energy	0.46	0.01	0.56	0.78
Immigration	0.52	0.01	0.47	0.63
Transportation	0.34	0.01	0.45	0.51
Crime & Law	0.30	0.03	0.42	0.62
Social Affairs	0.26	0.01	0.26	0.54
Housing & Community Development	0.44	0.00	0.44	0.57
Banking & Finance	0.22	0.02	0.32	0.59
Defence	0.38	0.02	0.48	0.65
Space, Science, Technology & Communications	0.28	0.01	0.36	0.64
Foreign Trade	0.19	0.01	0.17	0.54
International Affairs & Foreign Aid	0.31	0.04	0.29	0.62
Functioning Democracy	0.21	0.08	0.31	0.74
Public Lands & Water Management	0.16	0.01	0.13	0.42
Arts, Culture and Entertainment	0.29	0.04	0.51	0.76
Weather & Natural Disasters	0.35	0.01	0.50	0.71
Fires & Accidents	0.14	0.02	0.04	0.49
Sports & Recreation	0.38	0.02	0.63	0.78
Church & Religion	0.42	0.00	0.67	0.80

Table 6: Predicting Human Coding with Automated Coding using Marginal Probit, VRT and VTM Television (Belgium)

	Raw word counts		*Mutually-exclusive topics*	
	Pseudo-R2	**dfdx**	**Pseudo-R2**	**dfdx**
Macroeconomics	0.38	0.02	0.39	0.53
Civil Rights	0.19	0.07	0.18	0.57
Healthcare	0.42	0.04	0.52	0.70
Agriculture & Fisheries	0.44	0.02	0.55	0.76
Labour	0.42	0.03	0.46	0.54
Education	0.64	0.01	0.64	0.50
Environment	0.26	0.04	0.37	0.65
Energy	0.42	0.03	0.53	0.79
Immigration	0.61	0.01	0.57	0.65
Transportation	0.28	0.05	0.40	0.44
Crime & Legal Issues	0.37	0.16	0.50	0.74
Social Affairs	0.26	0.02	0.34	0.56
Housing & Community Development	0.28	0.02	0.32	0.40
Banking & Finance	0.17	0.07	0.28	0.56
Defence	0.29	0.05	0.42	0.64
Space, Science, Technology & Communications	0.31	0.03	0.48	0.81
Foreign Trade	0.22	0.01	0.26	0.61
International Affairs & Foreign Aid	0.33	0.08	0.40	0.66
Functioning Democracy	0.28	0.12	0.50	0.75
Public Lands & Water Management	0.09	0.01	0.15	0.36
Arts, Culture & Entertainment	0.36	0.06	0.63	0.79
Weather & Natural Disasters	0.54	0.07	0.74	0.89
Fires & Accidents	0.24	0.18	0.19	0.84
Sports & Recreation	0.64	0.30	0.88	0.96
Church & Religion	0.38	0.02	0.71	0.83

Table 7: Predicting Human Coding with Automated Coding using Marginal Probit, State of the Union

Topic Name	**Correlation**	**Coefficient**	R^2
Macroeconomics	0.80	0.74	0.64
Civil Rights, Minority Issues and Civil Liberties	0.40	0.51	0.16
Healthcare	0.84	0.93	0.71
Agriculture	0.89	0.93	0.79
Labor, Employment and Immigration	0.78	0.75	0.61
Education	0.91	0.70	0.83
Environment	0.61	1.20	0.37
Energy	0.93	1.38	0.87
Transportation	0.53	0.45	0.28
Law, Crime and Family Issues	0.64	1.06	0.41
Social Welfare	0.14	0.43	0.02
Community Development and Housing	0.65	1.00	0.42
Banking, Finance and Domestic Commerce	0.60	0.58	0.36
Defense	0.60	1.11	0.37
Space, Science, Technology and Communications	0.68	0.64	0.46
Foreign Trade	0.27	0.30	0.07
International Affairs and Foreign Aid	0.58	2.82	0.34
Government Operations	0.07	0.97	0.00
Public Lands and Water Management	0.26	0.41	0.07