# The Automated Coding of Policy Agendas: A Dictionary Based Approach

# The Automated Coding of Policy Agendas: A Dictionary Based Approach (v. 2.0.)

Julie Sevenans (University of Antwerp)
Quinn Albaugh (McGill University)
Tal Shahaf (Hebrew University of Jerusalem)
Stuart Soroka (McGill University)
Stefaan Walgrave (University of Antwerp)[1]

Paper prepared for the CAP Conference 2014, Konstanz.

**Abstract.** For the coding of political and media texts, the Policy Agendas community has mostly taken a human coding approach, or they have turned to automated machine learning methods. Last year, we proposed an alternative *dictionary-based* approach for the automated content analysis of texts (see Albaugh et al. 2013). We designed a first version of an English and a Dutch CAP-dictionary, and we validated the results of the codings against human coded documents. Although the results were not perfect yet – with respect to certain topic codes our dictionaries could certainly be improved – we showed that dictionaries may produce reliable, valid and comparable measures of policy and media agendas.

This year, we take the next step by doing three things. First, we further develop the Dutch and English dictionaries and try to replicate human coding results by making the dictionary assign single topic codes to single items. Second, we compare the dictionaries not only with human coded texts; we also validate them in a substantive manner. We expect individual Members of Parliament (MPs) to act most upon issues that they prioritize. Concretely, we test this by comparing MPs' dictionary-coded attention for issues with their committee membership. Third, we show that valuable insights can be gained from using the dictionaries in practice.

[1] The authors gratefully acknowledge support of the European Research Council (Advanced Grant 'INFOPOL', N° 295735) and of the Research Fund of the University of Antwerp (Grant N° 26827). Stefaan Walgrave (University of Antwerp) is principal investigator of the INFOPOL project, which has additional teams in Israel (led by Tamir Sheafer) and Canada (led by Stuart Soroka and Peter Loewen).

## Introduction

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 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[2]. 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, researchers on Policy Agendas 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.

Different other literatures explore the possibilities of "text as data" as well. Our own work—requiring a good deal of content analysis—is a good example. The INFOPOL project (http://www.infopol-project.org) investigates the information-processing behavior of political elites in Belgium, Canada and Israel; and it relies therefore not only on survey, interview and experimental data, but also on behavioral data: a content analysis of MPs' legislative behavior in Parliament. It is in this context that we started building dictionaries to automatically content analyze legislative data last year (Albaugh et al. 2013). Based on the Belgian and Canadian codebooks of the Comparative Agendas Project, we designed a first version of a Dutch and an English CAP-dictionary.

Given that many different and sophisticated supervised learning approaches exist for automated coding[3], why did we reconsider what seems to be a rather blunt, dictionary-

[2] The literatures are vast, but see Baumgartner and Jones 2005, Baumgartner et al. 2009, Baumgartner et al. 2011, Jones et al. 2003, Jones et al. 2009b, Jones and Baumgartner 2012.

[3] There is a wide literature about automated content analysis, both in general and related to the Policy Agendas Project. See Purpura and Hillard 2006; Hillard et al. 2007, 2008; 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 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—you get a frequency count based on a list of words, and that's all. Furthermore, dictionary-based systems may simply produce good results coding-wise. We think they are especially effective where the coding of legislative data is concerned, because policy topics are actually relatively easily identified using a finite set of keywords. Our own preliminary work suggests that economic policy does tend to include some very standard economic keywords; the same is true for labour policy, immigration, transportation; and so on. The end result is that sophisticated approaches may not produce systematically better codes than simple dictionary counts[4].

We concluded last year's paper with one major challenge: improving our dictionaries as to make them reliably assign a single topic code to a single text. Indeed, dictionary-based approaches have proven to be relatively good at identifying the various topics mentioned in a given text, but they have not often been used to conduct single-topic coding, and that is a key feature of the Policy Agendas coding scheme. That is the first goal of this paper. In the course of the last year, we worked on improving our dictionaries, trying to show *that dictionary coding can satisfactorily approximate single-topic human coding*.

This year we also want to validate our dictionaries in a more substantive manner. A thorough validation of an automatic coding approach goes beyond the pure comparison with human coding. We need to make sure that the dictionary produces meaningful results which are externally valid. That is our second concern here. Concretely, we check whether the dictionary coding can predict phenomena that one would expect to find in the real world. We try to demonstrate *that politicians' topic usage—as coded by the dictionaries—corresponds with their expected topic usage based on their positions in Parliament*.

Finally, we think that a promising line of research lies in the combination of content analyzed data with other types of datasets. In the framework of the INFOPOL project, the attitudinal (survey) and behavioral (content analysis) data gathered are complementary in the sense that they allow to compare what politicians *think* they do, with what they *actually* do. In this

[4] For a detailed comparison (and combination) of the two approaches, see Albaugh et al. 2014.

paper, we present some first, very preliminary results. Our third aim is as such to show *that valuable insights can be gained from using the dictionaries in practice*.

## 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 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[5]. Two years ago, 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. 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

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

with the Belgian Policy Agendas codebook, and some additional terms were added to the Dutch dictionary based on a consultation of the Belgian codes and inspired by the reading of different types of Belgian texts, such as newspaper articles. 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.

After the completion of the first version of the English and Dutch dictionaries last year, both dictionaries were updated during a thorough revision. Additional terms were added to the English-language dictionary based on items that appeared in the Dutch-language dictionary, and vice versa. A close reading of the U.S. and U.K. Policy Agendas codebooks also led to the inclusion of several additional terms. While the Belgian dictionary is still based on the original Belgian Policy Agendas codebook[6], the English dictionary corresponds to the master codebook, which immediately solves any potential problems caused by differences between the Canadian, U.S. and U.K. codebooks.

## Corpus selection

As the dictionaries are developed to investigate parliamentary behavior in the first place (mainly in the framework of the INFOPOL project), we rely on parliamentary debates to test the dictionaries.

To test the Dutch-language dictionary, we use the debates of the weekly plenary meeting in the Belgian federal Parliament, from the most recent legislature (July 2010 – April 2014)[7]. We include every word said by a Dutch-speaking MP or minister during the plenary meeting[8]—and these meetings include parliamentary questions and interpellations, answers on those questions, actual debates, discussions about legislative proposals, and so on. Our unit of analysis is an individual 'speech' (N = 9,073), defined as a talk given by one person without being interrupted. As soon as another person is called upon to speak, a new speech starts.

[6] This is done to be able to compare the dictionary codings with the original manual codings, which are not yet recoded according to the master codebook. In the near future, the Belgian dictionary can easily be adapted to the master codebook: we only need to move very few terms into different major topic codes.

[7] Those are retrieved from the official website of the Belgian federal Parliament: www.dekamer.be. The integral minutes of the plenary meeting are used.

[8] Except for the President of the Parliament, who conducts the meeting.

For the English-language dictionary, we draw on data from the U.K, for pragmatic reasons: while there is an extensive dataset on Question Period in Canada (Penner et al. 2006, Soroka et al. 2009, Young and Bélanger 2010), this dataset does not include any text relating to the questions. This leaves no text for a dictionary to code. As a result, we base our validation of the English-language dictionary against human coding on the Prime Minister's Question Time data from 1997-2008 posted on the U.K. Policy Agendas project website (N = 9,061).

## Analyses

In order to maintain comparability with most Policy Agendas datasets, we had to devise a way to assign one and only code to each speech using the dictionary word counts. To do this, we first selected the topic most frequently mentioned within each text[9]. This can only be done when there is a clear winner among the topics. That is, one topic is clearly mentioned more frequently than the others. Second, there are also texts that contain ties, meaning that two or more topics are equally present in the text. To break these ties, we took the topic first mentioned within each text. This is a fairly intuitive, but very reliable measure. Third, there are texts in which not any word of the dictionary is mentioned: for these texts we cannot attribute a code at all.

Based on the codings, we will now validate the dictionary approach in two ways. First, we will validate the Dutch-language and English-language dictionaries against human coding. Next, we do a substantive validation on the Belgian data only.

### Validation against human coding

We start with the validation of the Dutch dictionary. For the purpose of this paper, 500 randomly selected parliamentary speeches were coded by a human coder. This is a relatively low number of speeches, especially because some issues are rarely discussed in the Belgian federal Parliament. It leads to a problematically low N of manually coded speeches for certain major topic codes. In the future, we plan to have more speeches manually coded to

[9] Note that we ignore the fact that some topics have more keywords than others. In the English-language dictionary, for example, the number of keywords ranges from 28 to 143. It may be that these topics are more like to produce pluarality-winning totals. However, it seems more probable that there are topics that simply need fewer words to be identified. In the Dutch dictionary, for instance, the topic 'Immigration' counts only 35 words but is captured very well (see below); while the topic 'Human rights', with 102 words, performs rather poor.

conduct a more elaborate validation of the dictionary. For now, the 500 speeches are sufficient to give us a first indication of how well the dictionary performs against human coded texts.

Let us consider first the uncoded cases. Unlike, for example, parliamentary questions, that always deal with concrete policy issues, individual speeches in Parliament are not always about a policy domain. Sometimes an MP simply urges the other parliamentarians to remain silent while someone is speaking; or the so-called 'rapporteur' of a debate that has taken place during a committee meeting refers to the written report (standard procedures); and so on. In those instances, it is not desirable that a concrete major topic code is attributed to the speech; and even a human coder would not be able to do so. Table 1 shows this in numbers. It displays how many speeches could not be given a topic code, thereby comparing human and dictionary codes.

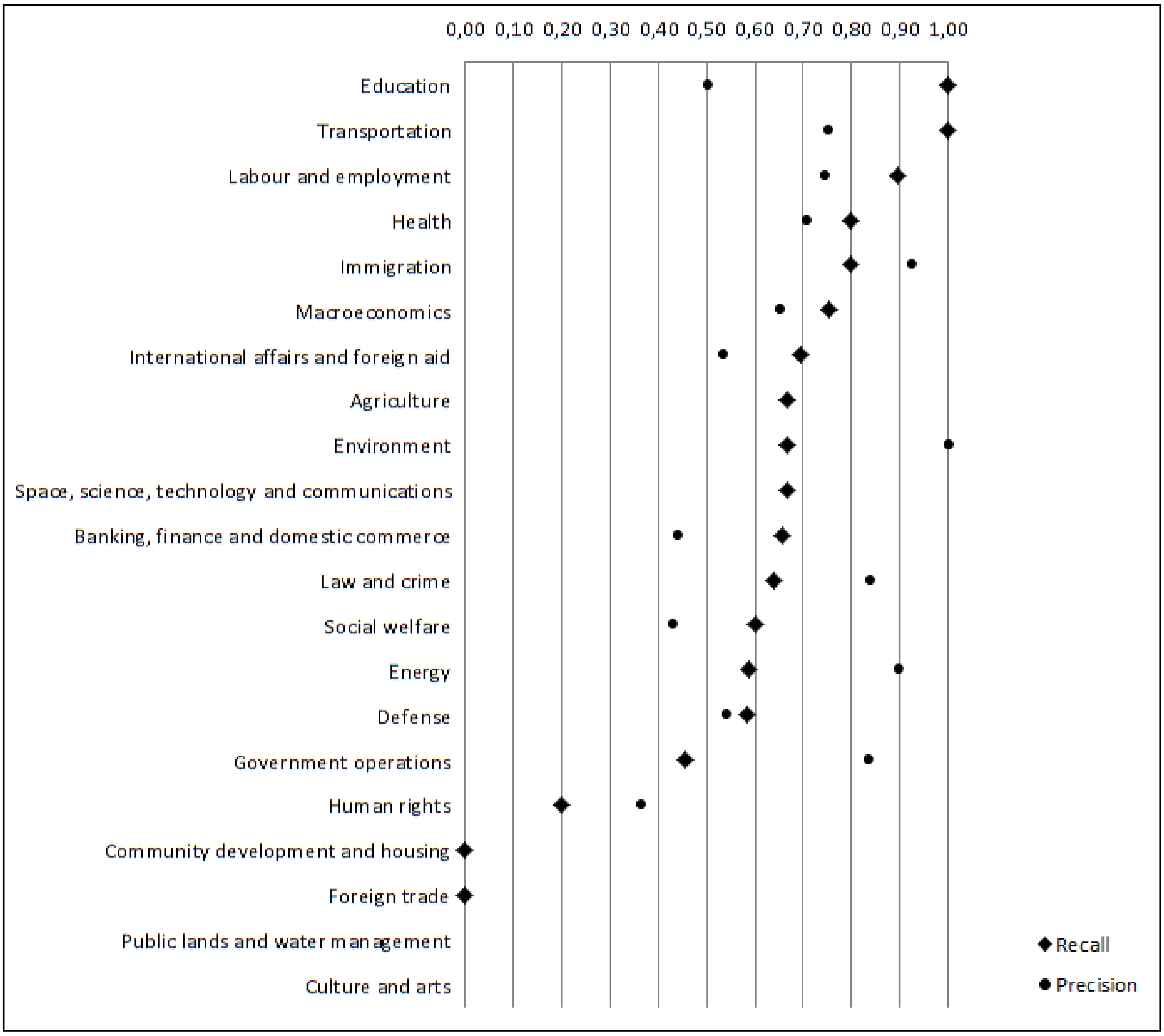
Table 1. Overview of not-coded speeches

| | **Coded by dictionary** | **Not coded by dictionary** | **Total** |
|---|---|---|---|
| **Coded by human coder** | 359 | 25 | 384 |
| **Not coded by human coder** | 11 | 105 | 116 |
| **Total** | 373 | 127 | 500 |

As we can see, most texts are either coded (359 speeches) or not-coded (105 speeches) by both the human coder and the dictionary. Only 25 speeches were manually coded but did not contain any word from the dictionary. As a diagnostic, we checked to see which kinds of questions—based on human-coded topics—were not coded with the dictionary. The 25 speeches are relatively evenly distributed over different topic codes (results not shown here) and there is not one topic that appears to be more problematic than the other topics in this sense. Furthermore, 11 speeches were coded by the dictionary, but not by the human coder. In total, there is a difference between human and dictionary in only 36 out of 500 speeches (7% of the data), which we think is acceptable. Let us now look at how well the dictionary-based coding holds up against the human coding in the instances where a topic could be assigned by the human coder.

Often used measures to validate automated coding against human coding are 'recall' and 'precision' (see Grimmer and Stewart 2013). 'Recall' refers to the number of documents correctly classified in topic *k*, divided by the total number of documents that the human coder classifies in topic *k*. It indicates thus what the chance is that the dictionary identifies the topic code that a human coder has assigned to a text. The 'precision' statistic is the number of documents correctly classified in topic *k*, divided by the total number of documents that the dictionary classifies in topic *k*. It gives an estimation of the chance that the topic code assigned to the text by the dictionary, has also been given by the human coder. The recall and precision for the 21 CAP major topic codes are visualized in Figure 1 and fully shown in Table 4 (in appendix).

Figure 1. Recall and precision for Belgian debates



We can see that the performance of the dictionary differs considerably by topic. Some topics perform really well: for Health, Labour, Immigration and Transportation, both recall and precision are higher than .70. 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 (see also Soroka 2014); however, codes that are measuring the same underlying phenomena should correspond fairly well to one another. For those four issues, this appears to be actually the case. Most other topics do moderately well, with recall and precision varying between .50 and .85. A few topics are still problematic, in particular 'Human rights' and 'Foreign trade' with very poor recall and precision. We will need to revise these topics in a next version of the dictionaries.

As explained above, the N (human coded) is very low for some issues; this is especially the case for those issues that are actually under the jurisdiction of the Belgian regional parliaments and that are thus almost never discussed in the federal Parliament, such as 'Agriculture', 'Education', and 'Community development and housing'. Two topics, 'Public lands and water management' and 'Culture and arts' do not even exist in the dataset. We cannot properly validate these topics here, but since the rest of the paper also uses data of the federal Parliament only, it does not pose a problem here. When applying the dictionaries in other contexts (e.g. on data of the regional parliaments or on media data) we will have to make sure that we do a proper validation test of these categories as well.

We can run similar tests on English language Prime Minister's Questions from the U.K. All questions in the dataset (N = 9,061) are coded by at least one human coder[10]. The questions are largely comparable to the speeches in Belgium (they are asked by one MP in Parliament without being interrupted), with one important difference: whereas not every Belgian speech was about an issue—some speeches were purely procedural, see above—the concrete U.K. questions are always about a topic. We thus try to code every item with the dictionary as well. Yet the English-language dictionary could not identify a topic for every single question. The results for uncoded items are displayed in Table 2.

[10] The UK Policy Agendas team had at least one coder read every piece. For much of the time period, they used two human coders, and for the rest they used supervised learning to train a second coder. If there was conflict, the project team leaders intervened. This is based on personal correspondence with Shaun Bevan. In subsequent analyses, it may be useful to subset the data using two human coders; however, our intent is to have human-coded debates with full text from the current Canadian parliament.

Table 2. Overview of not-coded questions

| | |
|---|---|
| **Dictionary code given** | 7,080 |
| **No dictionary code given** | 1,981 |
| **Total** | 9,061 |

In total, 1,981 questions (or 22% of all questions) did not contain any word from the dictionary and could therefore not be coded. The non-coded questions are relatively well spread across all different topics, with the exception of the topic ‘Government operations’. The clear majority (41 per cent) of questions not coded fell under this topic. 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.

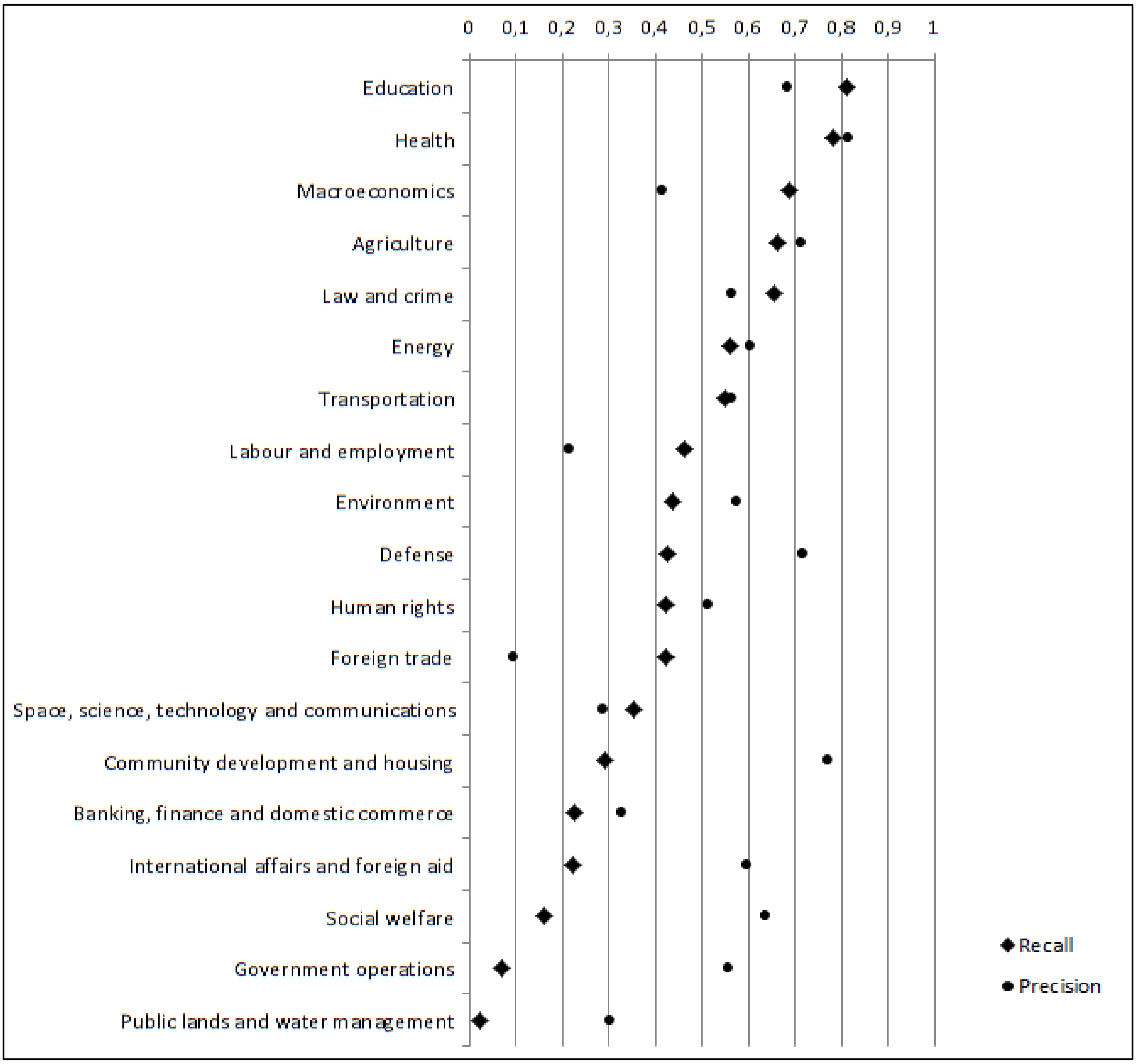
For the questions to which the dictionary could assign a winning topic code (either by taking the one most often mentioned topic, or by breaking the tie as explained above), we calculated recall and precision measures, like for the Belgian data. Figure 2 graphically shows the results; the numbers can be found in Table 5 (see appendix).

Similarly to the results of the Dutch dictionary, the recall and precision scores of the English dictionary differ a lot across topics. Health is the best performing topic with recall and precision of .78 and .81 respectively—in the Belgian case as well it was one of the topics that was best captured.

Overall, the mean recall and precision are higher for the Dutch speeches (recall 0.61; precision 0.60) than for the U.K. Prime Minister’s Questions (recall 0.43; precision 0.52). Note that caution is required when comparing the results of the Dutch and English dictionaries. With only 500 Dutch speeches coded manually, the test of the Dutch dictionary is a lot less robust than the English-language test. Furthermore, the type of texts coded is not identical, which may cause differences as well. For example, the mean word count of the Prime Minister’s Questions is 85, while the average Belgian speech consists of 375 words.

Since dictionaries seem likely to perform better when there is more text available, the difference in performance between the two dictionaries is (partly) to be expected.

Figure 2. Recall and precision for U.K. Prime Minister's Questions



The remarkably lower average recall score for the English questions seems to be mainly the consequence of a few really poorly performing topics (e.g. 'Social welfare', 'Banking, finance and domestic commerce', and 'International affairs and foreign aid') with recall below .30 for the English dictionary while it is above .60 for the Dutch dictionary. In the English-language dictionary, those specific topics certainly need improvement. On the other hand, topics such as 'Human rights' and 'Foreign trade' do somewhat better here than for the Dutch coding.

We can conclude for now that for both dictionaries to better match human coding, some topics need to be revised again[11]. In addition to the validation against human coding, scholars are encouraged to provide external validity to the data, by testing whether automatically coded data can predict relationships that one expects to exist in the real world. We now proceed with this more substantive validation, for the Dutch-language dictionary only.

## Substantive validation

Our substantive validation relies on the dataset of the Belgian speeches and consists of two tests: one based on MPs' committee membership, and another one based on issue diversity.

In order to conduct the substantive tests, the full dataset (N = 9,073) was coded. Of this total sample, 2,329 speeches were left uncoded (26%)—relatively speaking, the total dataset contains not more uncoded speeches than the subsample of 500 discussed above. For the other 6,744 speeches, we determined the single 'winning' topic code, either by taking the most often mentioned topic, or by 'breaking the tie' as explained above. Descriptive statistics are shown in Table 6 (see appendix). The Dutch-speaking members of the Belgian federal Parliament appear to talk most about 1) Law and crime, 2) Macroeconomics, and 3) Banking, finance and domestic commerce. Next, we aggregated the speeches on the level of the individual politician (N = 97), by calculating the average attention each MP or minister pays to each of the 21 different policy topic codes. These data were then merged with biographical data about the politicians, including their sex, function, committee membership, and so on.

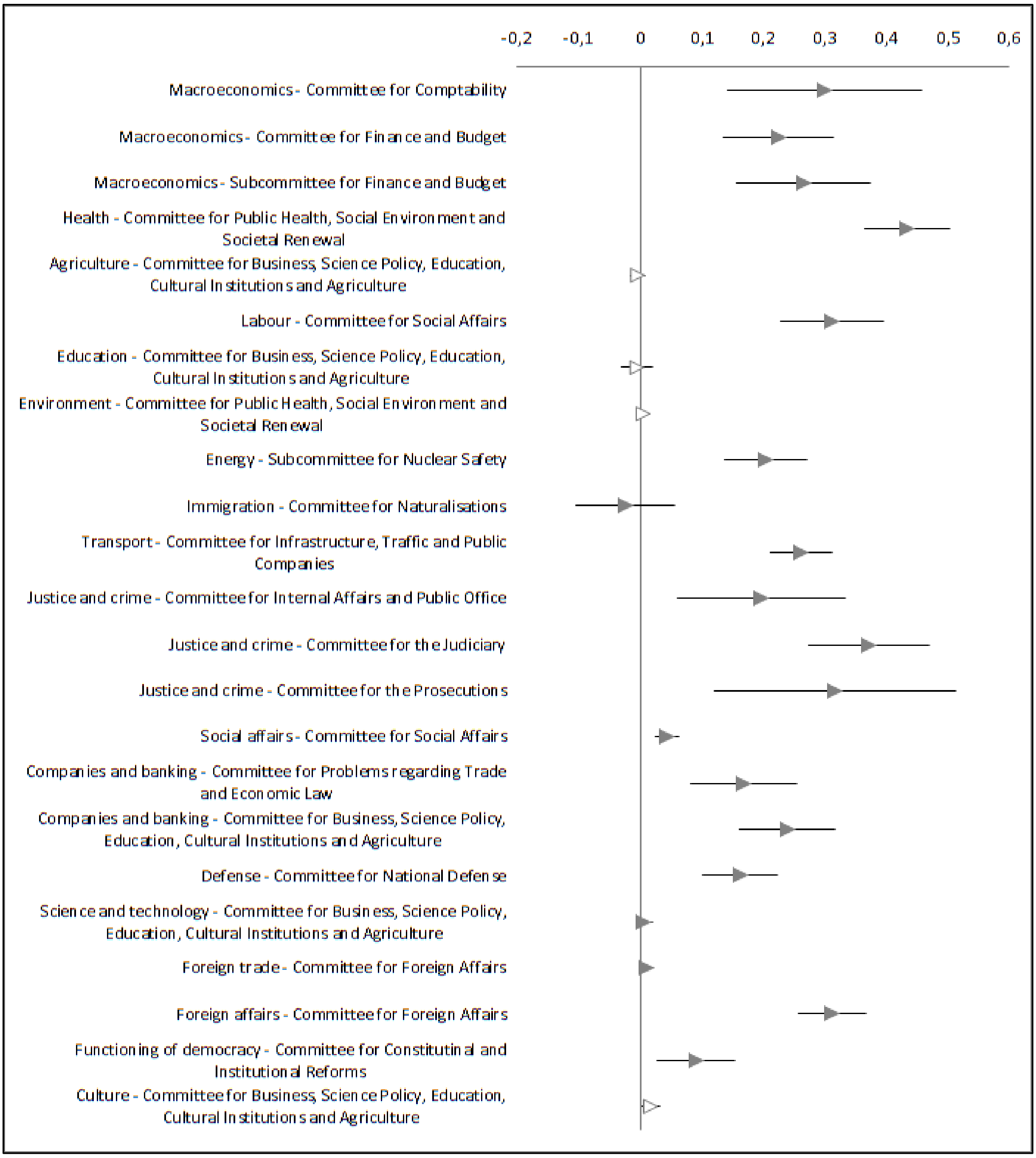
### *Committee membership*

In Parliament, we expect that members of a committee allocate more attention to issues that fall under the jurisdiction of their committee than other MPs who are not member of the committee (argument also made by Grimmer 2010; substantive validation method used by Grimmer and Stewart 2013). Note that we do not expect to find significant differences for every committee or every issue: for example, when an issue is very salient, many MPs will talk about it all the time, and we might expect committee membership to matter less.

[11] If the goal is not to directly match human coding but to allow for multiple topics, then the current dictionaries may be less problematic (Albaugh et al. 2014).

Rather, the idea here is to check whether we can find this pattern in general, for at least some issues, based on the dictionary-coded data. For this test, we only include MPs (N = 87).

Figure 3. Attention for issues in committees – difference between members and non-members (N = 87)



*Notes*. The triangles represent the (Average attention for an the, by the members of the committee) – (Average attention for the issue, by non-members of the committee). The lines are 95% confidence intervals. The white triangles indicate that the issue is formally a regional competence, and not a federal competence, explaining why federal MPs pay almost no attention to these issues.

The comparison between committee members' attention for the issue under the jurisdiction of the committee, and other MPs' attention, is graphically displayed in Figure 3. A table with the exact numbers and T-tests can be consulted in the appendix: see Table 7. The triangles represent the difference between members' and non-members' attention respectively. The grey-colored triangles represent issues that are a federal competence; the white-colored triangles represent regional competences. The lines are 95% confidence intervals.

Almost all grey-colored triangles lie on the right side of the full line, which means that the attention allocation patterns that we expected to find, are present in the data. Members of a committee pay, in many instances, 10% to 40% more attention to the related issue than non-members. The largest difference is found in the Committee for Public Health, where the members, on average, talk about health care in 43% more of their speeches than other MPs.

There are three federal issues for which we do not find similar effects: 'Immigration', 'Science and technology', and 'Foreign affairs'. Note that there is no clear link between how well an issue performs on the validation against human coding and the substantive validation: for the issue of immigration, for example, we established recall and precision scores of .80 and .90, and the fact that we do not find a difference in attention here probably means that members of the Committee for Naturalisations simply do not pay more attention to immigration than other MPs. And the results for 'Science and Technology' and 'Foreign Trade' may be due to the fact that these topics simply get very little attention in Parliament in general (see Table 6 in appendix). For the regional competences as well we do not find strong effects, which is not surprising.

Note that the varying sizes of the lines representing the 95% confidence intervals are mainly explained by the size of the respective committees. The smaller the committee, the larger the confidence interval.

### *Issue diversity*

For our second test of the predictive capacity of the dictionary codings, we make use of the concept of issue diversity. Maximal issue diversity is reached when every speech of a politician is about the same topic. Minimal diversity means that the politician distributes his talks equally across all 21 policy issues under study.

In order to measure issue diversity per politician, we calculate the 'Normalized Herfindahl index (H*)'[12]. The formula for this index is as follows:

$$H = \sum_{i=1}^{N} s_i^2$$

$$H^* = \frac{(H - 1/N)}{1 - 1/N}$$

The Normalized Herfindahl index ranges from 0 to 1, with 0 indicating maximal issue diversity and 1 indicating minimal issue diversity (or maximal issue concentration).

With respect to our dataset, we expect issue diversity to vary across political functions. Concretely, we anticipate ministers and committee chairs to be least issue-diverse. Ministers have one cabinet portfolio and in Parliament, they normally only get questions about this portfolio. A committee as well is by definition a very specialized institution and the chair of such a committee can therefore be expected to  be relatively focused on the issue of his committee. On the contrary, MPs who are leaders of a party or a faction have a very general profile and we foresee that they act on a very diverse range of issues in Parliament.

Figure 4. Issue diversity according to function

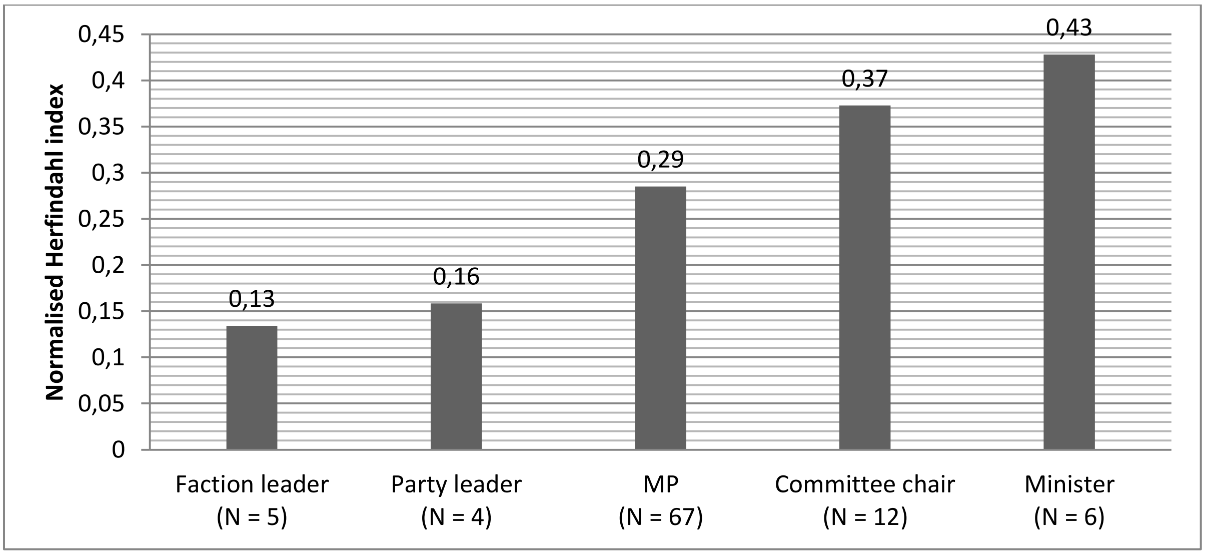


Figure 4 presents the results. The graph clearly shows that some functions have a less diverse issue profile than others. Ministers have, on average, the highest Normalized

[12] Note that we recognize the weaknesses of this measure, but here we just use it for diagnostic tests.

Herfindahl score (of 0,43), indicating that they are more focused (less issue-diverse) than politicians with another function. Faction leaders display the highest diversity (mean of Normalized Herfindahl index is 0.13). 'Normal' MPs (excluding committee chairs, faction and party leaders) occupy the middle position regarding diversity. The dictionary codings confirm the pattern that we expected, and the difference between the political functions is significant (ANOVA test: F = 2.87, $p < 0.05$).

We can conclude that, at least for the Belgian data, the substantive validation proves that the dictionary produces codings that make sense in the real world.

**Promising tracks of research**

Finally, we want to do a very first exploration of the possibilities offered by combining the richness of the content analyzed texts with other types of data. In the framework of the INFOPOL project, for example, we conducted a large survey with politicians in Belgium, Canada and Israel. In Belgium, federal Dutch-speaking MPs, ministers and party leaders were surveyed. The survey, administered on iPads, is part of a series of face-to-face elite interviews conducted between June and November 2013. In total, 87 out of 100 Belgian politicians participated, leading to a response rate of 87% which is exceptionally high for elite research. Each interview took approximately one hour and was scheduled beforehand. Most interviews took place in the MP's office in Parliament or in his/her home town.

The combination of the dictionary-coded behavioral data with the survey data allows us to compare what politicians *think* or *say* they do, with what they *actually* do. For example, we have a measure of politicians real, behavioral issue diversity in Parliament (as calculated above with the Normalized Herfindahl index), and we also personally asked them to estimate the degree to which they are issue-diverse, by the following question:

> "*Some politicians specialize in one or two policy areas, while others prefer to speak and act upon a wide range of issues from different policy areas. Where would you place yourself on a scale from 0 (I focus on one issue) to 10 (I focus on a broad range of issues)?*"

Let us calculate the correlation between politicians' self-reported issue diversity, and their objective, observed diversity. Note that we do not expect the two variables to perfectly correlate—they partly measure a different thing. But we expect there to be some overlap. The correlation between the Normalized Herfindahl index and the self-reported issue diversity is -0,35 (p = 0,001) (see Table 3). Politicians who say that they focus on many different issues, to a certain extent tend do so, and this relationship is significant. This is reassuring—it again validates our dictionary substantively.

The goal of the INFOPOL project is to study the information-processing behavior of political elites. Hence, one question battery in the survey gauges politicians' preference for different types of information. One of the items, related to information-processing and time horizons, is formulated as follows:

> "*Information can take different forms, and it can deal with different topics. Please indicate your preference on the following scale: Information that is useful in the short term (0) – Information that is useful in the long term (100).*"

Regarding the relationship between the issue diversity of politicians and the information preferences of these politicians, the INFOPOL project hypothesizes that specialists—who are very much focused on one or a few issues—have a preference for information that can be used in the long term. They thoroughly build a dossier on the topic, based on in-depth information. Generalists, on the other hand—politicians dealing with a diverse range of topics—are expected to prioritize information that they can use on the short term. They are rather the strategic type of politicians: they wait and see what the news of the day is and respond immediately.

To test this hypothesis, we can look at the relationship between issue diversity (both self-reported and observed) and the preference for short or long term information. The correlations are displayed in Table 3.

Table 3. Correlation between issue diversity (self-reported and observed) and information preference (N = 81)

| | **Self-reported issue diversity** | **Normalized Herfindahl index** | **Preference for long term info** |
|---|---|---|---|
| **Self-reported issue diversity** | 1,00 | - | - |
| **Normalized Herfindahl index** | -0,35** | 1,00 | - |
| **Preference for long term info** | -0,09 | 0,38*** | 1,00 |

*Note*. *** p < 0.001; ** p < 0.01; * p < 0.05

The correlation between the Normalized Herfindahl index, calculated based on the dictionary codings, and the preference for long term information is strong and significant (r = 0,38; p = 0,000). Politicians who talk about only one thing in Parliament, prefer information that can be used on the long term; while politicians talking about a broad range of issues rather like short term information. The hypothesis is confirmed. However, we do not find the same when using politicians' self-reported degree of issue diversity (r = -0,09, p = 0.42). The relationship goes in the expected direction but is not at all significant. A possible interpretation is that the survey data are subject to error—politicians cannot perfectly estimate their behavior and the interpretation of a survey question is always subjective—whereas the coding data offer a better way to objectively compare politicians.

Of course, this is only one, small, preliminary analysis and further exploration of the data is needed. But it gives us confidence that the data coded by the dictionaries offer many possibilities, and that valuable insights can be gained from using them in practice.

## Discussion and future steps

Overall, the results of the current paper are mixed. While our dictionaries improved in the course of the last year—topic codes that were problematic last year, now had acceptable recall and precision—the match with human coded texts is still not perfect. Note that this should not always be the primary goal: human coders make mistakes as well. And regarding reliability, for instance, we can be confident that dictionaries are extremely consistent. But if the goal is to make the dictionaries replicate manual coding, most issues need at least another round of revision.

However, while the dictionaries cannot (yet) perfectly approximate single-issue human coding, we already see real potential in the dictionaries as they are now. The results of the substantive validation of the Dutch dictionary were highly promising. We of course took advantage of what we know the dictionary is good at: identifying the relative importance of topics across many different texts (aggregate of all speeches by an MP in Parliament). When calculating measures like, for example, issue diversity, it is not so very important that each topic code works perfectly: as long as it captures most issues well, and as long as the weaknesses of the dictionary apply to each single text coded, the diversity score is reliable and comparable across MPs. In that sense, the data that we currently produce are readily usable, we think. In the near future, we intend to further explore the possibilities of the dictionary-coded data.

Finally, we would like to mention that we see real potential for the development of multi-lingual versions of the topic dictionaries. Towards that end, we have currently started building a Hebrew dictionary. Unfortunately, we could not include the results in this paper due to delays on several fronts—there were some problems with the data processing and the running of the software due to the fundamentally different characters and structure of the Hebrew language—but we hope to have the first results soon. The ability to reliably identify topics in many different languages is really attractive and we welcome efforts to translate the dictionaries into other languages. Our dictionaries—in Dutch, English and (in the near future) Hebrew—will in any case be made available at lexicoder.com.

## References

Albaugh, Quinn, Julie Sevenans, Stuart N. Soroka, and Peter John Loewen. 2013. “The Automated Coding of Policy Agendas: A Dictionary-Based Approach", paper presented at the 6th annual Comparative Agendas Project (CAP) conference, Antwerp, June 27-29.

Albaugh, Quinn, Stuart N. Soroka, Jeroen Joly, Peter Loewen, Julie Sevenans and Stefaan Walgrave. 2014. “Comparing and Combining Machine Learning and Dictionary-Based Approaches to Topic Coding", paper presented at the 7th annual Comparative Agendas Project (CAP) conference, Konstanz, June 12-14.

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* 2(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* 21(3):267-297.

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." 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, Bryan D, 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 comparative analysis." *American Journal of Political Science* 53(4):855-873.

Jones, Bryan D, T Sulkin, and HA Larsen. 2003. “Policy Punctuations in American Political Institutions." *American Political Science Review* 97(1):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.

Penner, Erin, Kelly Blidook, and Stuart Soroka. 2006. “Legislative Priorities and Public Opinion: Representation of Partisan Agendas in the Canadian House of Commons.” *Journal of European Public Policy* 13(7): 1006–20.

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 15 debate in the European Union." *European Union Politics* 11(3):333-357.

Soroka, Stuart. 2014. “Reliability and Validity in Automated Content Analysis." In *Communication and Language Analysis in the Corporate World*, pages 352-363, Hershey, PA: CGI Global.

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.

Soroka, Stuart, Erin Penner, and Kelly Blidook. 2009. “Constituency Influence in Parliament.” *Canadian Journal of Political Science* 42(3): 563–91.

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 Éric Bélanger. 2008. “BQ in the House: The Nature of Sovereigntist Representation in the Canadian Parliament.” *Nationalism and Ethnic Politics* 14(4): 487–522.

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.

## Appendix

Table 4. Recall and precision for Belgian debates

| Topic name | Recall | Precision | N (human coding) | Regional competence |
|---|---|---|---|---|
| Macroeconomics | 0.76 | 0.65 | 49 | |
| Human rights | 0.20 | 0.36 | 20 | |
| Health | 0.80 | 0.71 | 15 | |
| Agriculture | 0.67 | 0.67 | 3 | Yes |
| Labour and employment | 0.90 | 0.74 | 29 | |
| Education | 1.00 | 0.50 | 2 | Yes |
| Environment | 0.67 | 1.00 | 9 | Yes |
| Energy | 0.59 | 0.89 | 29 | |
| Immigration | 0.80 | 0.92 | 15 | |
| Transportation | 1.00 | 0.75 | 12 | |
| Law and crime | 0.64 | 0.84 | 64 | |
| Social welfare | 0.60 | 0.43 | 5 | |
| Community development and housing | 0.00 | 0.00 | 2 | Yes |
| Banking, finance and domestic commerce | 0.66 | 0.44 | 32 | |
| Defense | 0.58 | 0.54 | 12 | |
| Space, science, technology and communications | 0.67 | 0.67 | 6 | |
| Foreign trade | 0.00 | 0.00 | 2 | |
| International affairs and foreign aid | 0.70 | 0.53 | 23 | |
| Government operations | 0.45 | 0.83 | 55 | |
| Public lands and water management | / | / | 0 | Yes |
| Culture and arts | / | / | 0 | Yes |

Table 5. Recall and precision for U.K. Prime Minister’s Questions

| Topic name | Recall | Precision | N (human coding) |
|---|---|---|---|
| Macroeconomics | 0.69 | 0.41 | 638 |
| Human rights | 0.42 | 0.51 | 329 |
| Health | 0.78 | 0.81 | 864 |
| Agriculture | 0.66 | 0.71 | 172 |
| Labour and employment | 0.46 | 0.21 | 274 |
| Education | 0.81 | 0.68 | 568 |
| Environment | 0.44 | 0.57 | 137 |
| Energy | 0.56 | 0.60 | 137 |
| Transportation | 0.55 | 0.56 | 336 |
| Law and crime | 0.66 | 0.56 | 778 |
| Social welfare | 0.16 | 0.63 | 415 |
| Community development and housing | 0.29 | 0.77 | 227 |
| Banking, finance and domestic commerce | 0.23 | 0.32 | 284 |
| Defense | 0.43 | 0.71 | 1,163 |
| Space, science, technology and communications | 0.35 | 0.28 | 88 |
| Foreign trade | 0.42 | 0.09 | 45 |
| International affairs and foreign aid | 0.22 | 0.59 | 894 |
| Government operations | 0.07 | 0.55 | 1,569 |
| Public lands and water management | 0.02 | 0.30 | 143 |

Table 6. Proportion of speeches per topic (N = 6,744) - Belgium

| **Topic** | **Mean** | **S.D.** |
|---|---|---|
| Macroeconomics | 0,1409 | 0,3479 |
| Human rights | 0,0328 | 0,1780 |
| Health | 0,0420 | 0,2005 |
| Agriculture | 0,0046 | 0,0676 |
| Labour and employment | 0,1016 | 0,3021 |
| Education | 0,0079 | 0,0883 |
| Environment | 0,0074 | 0,0858 |
| Energy | 0,0466 | 0,2107 |
| Immigration | 0,0586 | 0,2348 |
| Transportation | 0,0580 | 0,2337 |
| Law and crime | 0,1649 | 0,3711 |
| Social welfare | 0,0205 | 0,1416 |
| Community development and housing | 0,0034 | 0,0583 |
| Banking, finance and domestic commerce | 0,1272 | 0,3332 |
| Defense | 0,0348 | 0,1834 |
| Space, science, technology and communications | 0,0105 | 0,1021 |
| Foreign trade | 0,0018 | 0,0421 |
| International affairs and foreign aid | 0,0621 | 0,2414 |
| Government operations | 0,0703 | 0,2556 |
| Public lands and water management | 0,0013 | 0,0365 |
| Culture, arts | 0,0030 | 0,0544 |

Table 7. Committee membership test

| Issue – committee | N | Non-member | Member | Difference | t | |
|---|---|---|---|---|---|---|
| Macroeconomics - Committee for Comptability | 3 | 0,1115 | 0,4106 | 0,2992 | 3,75 | *** |
| Macroeconomics - Committee for Finance and Budget | 9 | 0,0985 | 0,3232 | 0,2247 | 4,95 | *** |
| Macroeconomics - Subcommittee for Finance and Budget | 6 | 0,1035 | 0,3688 | 0,2653 | 4,84 | *** |
| Health - Committee for Public Health, Social Environment and Societal Renewal | 11 | 0,0227 | 0,4567 | 0,4340 | 12,27 | *** |
| Agriculture - Committee for Business, Science Policy, Education, Cultural Institutions and Agriculture | 10 | 0,0063 | 0,0018 | -0,0044 | -0,73 | |
| Labour - Committee for Social Affairs | 11 | 0,0662 | 0,378 | 0,3118 | 7,45 | *** |
| Education - Committee for Business, Science Policy, Education, Cultural Institutions and Agriculture | 10 | 0,0111 | 0,0053 | -0,0057 | -0,46 | |
| Environment - Committee for Public Health, Social Environment and Societal Renewal | 11 | 0,0062 | 0,0107 | 0,0045 | 0,78 | |
| Energy - Subcommittee for Nuclear Safety | 7 | 0,0239 | 0,2281 | 0,2042 | 6,03 | *** |
| Immigration - Committee for Naturalisations | 8 | 0,0446 | 0,0202 | -0,0244 | -0,60 | |
| Transport - Committee for Infrastructure, Traffic and Public Companies | 10 | 0,0216 | 0,2821 | 0,2605 | 10,35 | *** |
| Justice and crime - Committee for Internal Affairs and Public Office | 9 | 0,1371 | 0,3336 | 0,1966 | 2,86 | ** |
| Justice and crime - Committee for the Judiciary | 12 | 0,1062 | 0,4771 | 0,3709 | 7,54 | *** |
| Justice and crime - Committee for the Prosecutions | 4 | 0,1429 | 0,4588 | 0,3160 | 3,20 | ** |
| Social affairs - Committee for Social Affairs | 11 | 0,014 | 0,0578 | 0,0437 | 4,35 | *** |
| Companies and banking - Committee for Problems regarding Trade and Economic Law | 10 | 0,099 | 0,2677 | 0,1687 | 3,89 | *** |
| Companies and banking - Committee for Business, Science Policy, Education, Cultural Institutions and Agriculture | 10 | 0,091 | 0,3298 | 0,2388 | 6,09 | *** |
| Defense - Committee for National Defense | 11 | 0,0225 | 0,1851 | 0,1625 | 5,24 | *** |
| Science and technology - Committee for Business, Science Policy, Education, Cultural Institutions and Agriculture | 10 | 0,0076 | 0,0136 | 0,0060 | 0,82 | |
| Foreign trade - Committee for Foreign Affairs | 9 | 0,0004 | 0,0107 | 0,0104 | 5,67 | *** |
| Foreign affairs - Committee for Foreign Affairs | 9 | 0,0386 | 0,3502 | 0,3116 | 11,07 | *** |
| Functioning of democracy - Committee for Constitutinal and Institutional Reforms | 9 | 0,0536 | 0,1444 | 0,0909 | 2,85 | ** |
| Culture - Committee for Business, Science Policy, Education, Cultural Institutions and Agriculture | 10 | 0,0038 | 0,0205 | 0,0168 | 2,30 | * |

*Note*. *** p < 0.001; ** p < 0.01; * p < 0.05