IIIS Discussion Paper

No.338 / July 2010

A new Database of Parliamentary Debates in Ireland, 1922--2008

Slava Mikhaylov
London School of Economics, Institute for International Integration Studies, Trinity College Dublin

Alexander Herzog
New York University

IIIS Discussion Paper No. 338

A new Database of Parliamentary Debates in Ireland, 1922--2008

Slava Mikhaylov
London School of Economics, Institute for International Integration Studies, Trinity College Dublin

Alexander Herzog
New York University

A new Database of Parliamentary Debates in Ireland, 1922–2008*

Alexander Herzog
New York University
alexander.herzog@nyu.edu

Slava Mikhaylov
London School of Economics
v.mikhaylov@lse.ac.uk

July 26, 2010

Abstract

We present a new database of parliamentary debates and written answers in Dáil Éireann for the entire time period from the third Dáil in 1922 to the thirtieth Dáil in 2008. This database was built from the Official Records of the Houses of the Oireachtas. Unlike its original version, our database integrates information about debates and information about deputies into a single database. This database therefore allows to search and retrieve contributions from individual deputies of the Dáil (Teachta Dála or TD) and to combine information about TDs' parties and constituencies with the history of political speeches and written answers. In addition, our database facilitates the application of content analysis software such as Wordscore (Laver, Benoit and Garry, 2003) or Wordfish (Slapin and Proksch, 2008) and makes it possible to estimate TDs' policy preferences from speeches. In this paper we document the structure of the database and how it was generated. We furthermore demonstrate how political debates can be used in social science research through a series of examples. These include an analysis of the policy agenda in all budget speeches from 1922 to today, the estimation of speakers' policy positions in the 2008 budget debate, and the estimation of ministers' policy positions in the 26th government in 2002.

Key Words: Parliamentary debates, policy point estimation, budget speeches, text analysis

*Note: This manuscript is still work in progress — comments are welcome and we urge you to wait until we declare a final version before citing it. This research was supported in part by the Irish Research Council for Humanities and the Social Sciences. We thank Patrick Honohan for invaluable comments at various stages of this project. We also thank the Institute for International Integration Studies (IIIS) at Trinity College Dublin for providing research facilities during this project. Authors' names are listed in alphabetical order. Authors have contributed equally to all work.

1 Introduction

Almost all political decisions and political opinions are, in one way or the other, expressed in written or spoken text. Leaders in history became famous for their ability to motivate masses with their speeches; parties publish policy programmes before elections in order to provide information about their policy objectives; parliamentary decisions are discussed and deliberated on the floor in order to exchange opinions; members of the executive in most political systems are legally obliged to provide written or verbal answers to questions from legislators; and citizens express their opinions about political events in internet blogs or in public online chats. Political texts and speeches are everywhere where people express their political preferences.

It is not until recently that social scientist have discovered the potential of analysing political texts to test theories of political behavior. One reason is that systematically processing texts to retrieve information is technically challenging. New approaches in computerised content analysis have greatly facilitated this task. Statistical techniques such as Wordscore (Benoit and Laver, 2003; Laver, Benoit and Garry, 2003) or Wordfish (Slapin and Proksch, 2008) now enable researchers to systematically compare documents with each other and to extract relevant information from them. Applied to party manifestos, for which most of these techniques have been developed, these methods can be used to evaluate the similarity or dissimilarity between manifestos, which then can be used to derive estimates of parties' policy preferences and their distance to each other.

One area of research that increasingly makes use of quantitative text methods are studies of legislative behaviour (Giannetti and Laver, 2005; Laver and Benoit, 2002; Monroe, Colaresi and Quinn, 2008; Proksch and Slapin, 2009*b*; Yu, Kaufmann and Diermeier, 2008;

Charbonneau, 2009; Galli, Grembi and Padovano, 2009; Imbeau, 2009; Hopkins and King, 2010; Quinn et al., 2010; Grimmer, 2010). Only a few parliaments in the world use roll-call votes (the recording of each legislator's decision in a floor vote) that allow to monitor individual members' behaviour. In all other cases, contributions to debates are the only outcome that can be observed from individual members. Using such debates for social science research, however, is often limited by data availability. Although most parliaments keep written records of parliamentary debates and often make such records electronically available, they are never published in formats that facilitate social science research. A significant amount of labour is usually required to collect, clean and organise parliamentary records before they can be used for statistical purposes, often requiring technical skills that many social scientists lack.

The purpose of this paper is to present a new database of parliamentary debates to overcome this hurdle. Our database contains all debates as well as questions and answers from the third to the thirtieth Dáil Éireann (1922–2009), covering almost a century of political discourse. These debates are organised in a way that allows users to search them by date, topics or speaker. More importantly, and lacking in the official records of parliamentary debates, we have identified all speakers and linked their debate contributions to the information on party affiliation and constituencies from the official members database. This enables researchers to retrieve member-specific speeches on particular topics or within a particular time frame, which is necessary to apply computerised content analysis software. Furthermore, all data can be retrieved and stored in formats that can be accessed using commonly used statistical software packages.

In addition to documenting this database, we also present three applications in which we make use of the new data (Section 3). In the first study, we analyse budget speeches

delivered by all finance ministers from 1922 to 2008 (Section 3.1) and show how the policy agenda and ministers' policy preferences have changed over time (Section 3.2). In the second application we compare contributions that were made on one particular topic: the 2008 budget debate (Section 3.3). Here we demonstrate how computerised content methods can be used to estimate members' policy preferences on a dimension that represents pro- versus anti-government attitudes. Finally, we estimate all contributions from members of the 26th government that formed as a coalition between Fianna Fáil and the Progressive Democrats in 2002. Here we estimate the policy positions of all cabinet ministers on a pro- versus anti-spending dimension and show that positions on this dimension are highly correlated with actual spending levels of each ministers' department (Section 3.4).

2 Overview of Database Content

Parliamentary debates in Dáil Éireann are collected by the Oireachtas Debates Office and published as the Official Record. The Debates Office records and transcribes all debates and makes them available in both printed and digital form. Electronic versions of debates are published as single HTML files on the Houses of Oireachtas website.[1] At the moment of writing, the official debates website contains 549,292 HTML files. The content of all these HTML files forms the data source for our database. It is obviously impossible to hand-code that much information. We therefore wrote a computer script that automated the processing of all files.[2] This script is able to find all debate contributions and the names of all speakers in each file. In addition, it retrieves the date as well as the topic of each debate.

[1] Official records are available at http://debates.oireachtas.ie/Main.aspx (last accessed on 17 December 2009). More detailed information about the Debates Office's work can be found at http://www.oireachtas.ie/viewdoc.asp?fn=/documents/Organisation/debatesoffice2.htm (last accessed on 17 December 2009).

[2] The computer script consists of multiple syntax files that were written in the computer language Python.

The officially published debate files contain the name of each speaker. These names, however, are "hard coded" into the HTML files and not linked to the information in the official members database. In addition, speaker names are not coded consistently, hence making it difficult to collect speeches from a particular deputy.[3] Our goal was to identify every single speaker name that appears in the Official Record and to integrate parliamentary speeches with information about deputies' party affiliation, constituency, age and profession from the official members database into a single database. We therefore used an automated record-linkage procedure to identify every speaker. [4]

The final database contains all debates and written answers from the first meeting of the 3rd Dáil on 9 September 1922 through the end of 2008, covering every Dáil that has met during this period. In total, the database contains 7,006 parliamentary sessions with a total of 440,223 individual contributions. Taking all debates together, our database contains 434,018,123 words with an average length of about 1,000 words per contribution. The data is organised in a way that facilitates the application of computerised content analysis software. Every row in the data set is one contribution with columns containing information on the following variables:[5]

[3]More recent parliamentary debates are made available in a dynamic framework on the Oireachtas' website. This new interface allows to retrieve speaker-specific information and to retrieve speeches from a single member. However, this only applies to debates since 2007.

[4]Record-linkage is a common technique that is used to link entries from two databases that share the same content but differ in how entries are coded. The basic idea of this procedure is to compare every entry from one database (in our case, the complete list of all speaker names) with every entry from the second database (in our case, the official members database), using some pre-defined algorithm to determine which two entries are most similar to each other. Different record-linkage algorithms have been developed and, after comparing several algorithms, we found the "longest common sub-string" procedure to work particularly well with our data. (See Christen (2006) for an overview and comparison of different record-linkage procedures.) The computer code we applied comes from *Febrl*, a Python environment that was developed by the ANU Data Mining Group at the Australian National University (http://datamining.anu.edu.au/linkage). We thank Holger Döring for making us aware of *Febrl*.

[5]The data is actually stored in a relational database in order to avoid recording redundant information. However, almost all statistical software packages require data that are organised in spreadsheet format. That is, with rows containing observations and columns containing information on variables. Here we describe the format of the spreadsheet files as these are the files that users will download and use for their analyses.

- the first name and surname of the speaker,
- the speaker's party affiliation and constituency,
- the Dáil and volume number from the official records,
- the day, month and year of the debate,
- the topic of the debate,
- a number representing the place of the contribution in the order of all contributions that were made during a particular debate.

The following page shows a small sample of the data set for speeches from finance minister Brian Lenihan.

A sample of the data with speeches from finance minister Brian Lenihan

First name	Surname	Party	Constituency	Dail	Volume	Year	Month	Day	Topic	Order	Contribution
Brian	Lenihan	FF	Dublin West	30	663	2008	10	14	Budget Debate 2009	2	We find ourselves in one of the most difficult and uncertain times in living memory. Turmoil in the financial markets and steep increases in commodity prices have put enormous pressures on economies throughout the world. Here at home we face the most challenging fiscal and economic position in a generation. [...]
Brian	Lenihan	FF	Dublin West	30	663	2008	10	14	Written Answer – Public Sector Pay	2	The Government continues to be committed to safeguarding the taxpayer's money by ensuring that delivery of public services in the most efficient and effective manner. The recent Review and Transitional Agreement under Towards 2016, which has yet to be ratified, recognises that the Public Service must review continuously its systems, processes and procedures, to ensure that it is responsive and efficient and that it provides high quality, value for money services. [...]
Brian	Lenihan	FF	Dublin West	30	663	2008	10	14	Written Answer – National Development Plan	2	Funding under the NDP (2007-2013) in 2009 and subsequent years is an issue that will be addressed in the 2009 Budget being presented to the Dáil today.
...	...	...	...	...	...	...	...	...	...	...	
Brian	Lenihan	FF	Dublin West	30	665	2008	10	30	Adjournment Debate. - Financial Institutions Support Scheme	2	Deputy Burton has raised the issue of the serious financial consequences to the State arising from the guarantee scheme to banks and [628] credit institutions. In the course of her contribution she raised a number of issues which do not arise within the terms of the Adjournment matter. However, I will be pleased to deal with these by way of a reply to a parliamentary question. [...]
...	...	...	...	...	...	...	...	...	...	...	

By default (that is, in the data sets that we make available online) contributions are sorted by members, making it very easy to apply computerised content analysis software that requires text being organised by individuals. Contributions can also be sorted by any other column, e.g. by parties which would allow to compare contributions from members of one party to that of another. In addition, we provide a variable "order" that indicates the order in which contributions were made during a debate, hence making it possible to reconstruct the precise sequence of a debate or question and answer session.

3 Analysing the Content of Parliamentary Debates

In the previous section, we have explained the structure of the database. In the following three sections we demonstrate how the data can be used for social science research. We do this by demonstrating three different applications. In the first application, we analyse the budget speeches of all finance ministers from 1922 to 2008. The resulting data set resembles a time series because we observe a single speech per year but over a long time period. [6] Analysing this data, we show how policy agenda and ministers' fiscal preferences have changed over time. In the second application, we construct a data set that resembles a cross-sectional analysis because we retrieve all speeches from one particular year and on one particular topic from our database: the 2008 budget debate. This data structure enables us to estimate the policy positions of all speakers who contributed to the budget debate and to compare how similar or dissimilar their preferences were. We find that policy positions are clustered into two groups: the government and the opposition; but we also find considerable variation within each group. Finally, we take all contributions that were made

[6] An exception are emergency budgets which obviously increase the number of budget speeches in a year.

during the term of one government and use the data to estimate the policy positions of all cabinet members on a dimension representing pro- versus anti-spending. We demonstrate the validity of estimated policy positions by comparing them against actual spending levels of each cabinet ministers' department and show that the two measures are almost perfectly correlated with each other.

3.1 The Content of Budget Speeches in Historical Perspective

The quantitative analysis of text is primarily based on the proposition that preference profiles of speakers can be constructed from their word frequencies (Baayen, 2001; Bybee, 2001). This makes word frequencies the most important data input to almost all existing methods of text analysis. Word frequencies can be easily visualised as *word clouds*. These word clouds show the most frequently used words in a text with font size being proportional to frequency of appearance. This method is popular with political pundits and has been popularised over the last presidential election in the US, where it was used to "visualize some of the most pressing issues that the presidential candidates would like to imprint upon voters' minds" (Yao, 2008). This method has also been used by mainstream media to analyse a wide variety of policies. For example, Obama's statements in the health care debate (Gavin, 2009); verbal duelling of Obama and Cheney on national security (Condon, 2009); Obama's inauguration speech (Day, 2009); and policy priorities of David Cameron (Williams, 2008). The Washington Post has now started a dedicated project that analyses the frequency of words in all speeches made by President Obama using word clouds.[7]

Here we show how word frequencies can be used to analyse speeches made by Irish Ministers for Finance. We have extracted the budget speeches of all finance ministers from

[7] http://projects.washingtonpost.com/obama-speeches/ (Accessed 2010-01-27)

our database, the first being Cosgrave's speech in April 1923, and the latest being Lenihan's speech in October 2008. In total, there are 90 speeches given by 23 different finance ministers for whom we have generated word clouds as shown in Figure 1.

[∗∗∗∗ FIGURE 1 ABOUT HERE ∗∗∗∗]

Each individual word cloud panel presents a snap shot into the preference profiles of individual ministers. With taxation being the key instrument of fiscal policy it is not surprising that the word "tax" is on average the most frequently used word across all Ministers for Finance. We can also discern that the frequency of references to "government" has been uneven over time with a relatively high usage in the 1960s to 1980s and then a subsequent decline (apart from Quinn's tenure) until Cowen and particularly Lenihan's speeches.

What is more clearly evident is the change in the number of unique words used by different ministers. This reflects the fact that some budget speeches were very short, while others were long and covered many distinct topics. The easiest example is to compare speeches by two consecutive ministers, e.g. for Cowen and Lenihan. Word clouds reflect the sheer multitude of problems facing the country that needed to be addressed by Lenihan compared to the relatively "quieter" (on average) three budgets delivered by Cowen.

Overall, while being catchy, word clouds can only be used as easy first-cut visualisations of the data. One thing that becomes readily apparent from Figure 1 is that word clouds do not facilitate systematic comparison of documents and their content with each other. Pundits usually interpret word clouds simply by pointing out prominent features (i.e., the most high-frequency words) (e.g. Condon, 2009). Very often word clouds are presented for consumption with a suggestion to "[t]ake a look for yourself" (Gavin, 2009) or "[t]ake from this what you will" (Yao, 2008). Recent advances in text analysis allow us to go

further than simple punditry. In the next section we demonstrate how our data facilitates the application of text analysis techniques to answer more complex empirical questions without the ambiguity in interpretation that is inherent in word clouds.

3.2 Estimation of Finance Ministers' Policy Positions

Wordfish (Slapin and Proksch, 2008) is a method that combines Item Response Theory (e.g. Clinton, Jackman and Rivers, 2004) with text classification. Wordfish assumes that there is a latent policy dimension and that each author has a position on this dimension. Words are assumed to be distributed over this dimension such that $y_{ijt} \sim Poisson(\lambda_{ijt})$, where y_{ijt} is the count of word j in document i at time t. The functional form of the model is assumed to be

$$\lambda_{ijt} = \exp(\alpha_{it} + \psi_j + \beta_j \omega_{it})$$

where α_{it} are fixed effects to control for differences in the length of speeches and ψ_j are fixed effects to control for the fact that some words are used more often than other words in all documents. ω_{it} are the estimates of authors' position on the latent dimension and β_j are estimates of word-weights that are determined by how important specific words are in discriminating documents from each other. In this model each document is treated as a separate actor's position and all positions are estimated simultaneously. If a minister maintains a similar position from one budget speech to the next, it means that the minister used words with similar frequencies over time. At the same time, any movement detected by the model towards a position held by, for example, his predecessor, means that minister's word choice is now much closer to his predecessor than to his own word usage in the previous budget speech. The identification strategy for the model also sets the mean of all positions to 0

and the standard deviation to 1, thus allowing over time change in positions relative to the mean with the total variance of all positions over time fixed (Slapin and Proksch, 2008). Effectively this standardises the results and allows comparison of positions over time on a comparable scale.

Before including documents in the analysis, we have removed all numbers, punctuation marks, and stop words. In addition, we followed the advice in Proksch and Slapin (2009*a*) and deleted words that appear in less than 20% of all speeches. We do this in order to prevent words that are specific to a small time period (and hence only appear in a few speeches) to have a too large impact on discriminating speeches from each other.

Figure 2 shows the results of estimation, including a line representing fitted values from a linear regression.

[$****$ FIGURE 2 ABOUT HERE $****$]

The results in Figure 2 indicate a policy agenda shift over time. We can also observe that some ministers have similar preference profiles while others differ significantly. For example, Ahern and Reynolds are very similar in their profile but differ from a group consisting of Quinn, McCreevy, Cowen and Lenihan who are very close to each other. There also appears to be a dramatic shift in agenda between the tenures of Lynch and Haughey (and also during Taoiseach Lynch's delivery of the budget speech for the Minister for Finance Charles Haughey in 1970). Overall, it appears that topics covered in budget speeches develop in waves, with clear bands formed by, for example, Lenihan, Cowen, McCreevy and Quin; Ahern and Reynolds; MacSharry, Dukes, Bruton, Fitzgerald, O'Kennedy and Colley; R.Ryan, Colley, Lynch (for Haughey); MacEntee, McGilligan and Aiken; Blythe and MacEntee.

The shift in positions of individual Ministers observed in Figure 2 can be explained in terms of the changing economic environment in the country (thus capturing the policy agenda shift) or in terms of individual idiosyncrasies of Ministers. We consider the relationship between estimated policy positions of Minsters and three core economic indicators: unemployment, inflation, and per capita GDP growth rates. Figure 3 shows the three economic indicators, inflation (1923–2008), GDP growth (annual %; 1961–2008) and unemployment rate (1956–2008), over time. Figures 4, 5 and 6 show Ministers' estimated positions plotted against the three indicators.

[∗∗∗∗ FIGURE 4 ABOUT HERE ∗∗∗∗]

[∗∗∗∗ FIGURE 5 ABOUT HERE ∗∗∗∗]

[∗∗∗∗ FIGURE 6 ABOUT HERE ∗∗∗∗]

As expected, the results show that policy positions of some Ministers can be explained by the contemporaneous economic situation in the country. However, the fact that some of the Ministers are clear outliers highlights the effect of individual characteristics on policy-making. One of the avenues for research that arises from this exercise is to analyse the determinants of these individual idiosyncrasies, possibly looking at education, class, and previous ministerial career. We claim that this and related questions can now be easily investigated by researchers using our database.

3.3 Speakers' Policy Position in the 2008 Budget Debate

In the previous section, we used budget speeches from each year and compared them over time. In this section, we restrict the analysis to a single year but take multiple speeches

made on the same topic. More precisely, we estimate the preferences of all speakers who participated in the debate over the 2008 budget. We extract these speeches from the database by selecting all contributions with topic "Financial Resolution" in year 2007.[8] This leaves us with a total of 22 speakers from all five parties. Table 1 shows the speeches included in the analysis.

[∗∗∗∗ TABLE 1 ABOUT HERE ∗∗∗∗]

To estimate speakers' position we use Wordscore, a computer algorithm developed by Laver, Benoit and Garry (2003). In a similar application, Laver, Benoit and Garry (2003) have already demonstrated that Wordscore can be effectively used to derive estimates of TDs policy positions.[9] As in the example above, we pre-processed documents by removing all numbers and interjections.

Wordscore uses two documents with well known positions as reference texts. The positions of all other documents are then estimated by comparing them to these reference documents. The underlying idea is that a document that, in terms of word frequencies, is similar to a reference document was produced by an author with similar preferences. The selection of reference documents furthermore determines the (assumed) underlying dimension for which documents' positions are estimated. For example, using two opposing documents on climate change would scale documents on the underlying dimension "climate politics". It has also been shown that under certain assumptions the Wordscore algorithm is related to the Wordfish algorithm that we used in the previous section (Lowe, 2008). Kluver (2009), furthermore, showed that the results of both procedures are highly correlated.

[8]The debates for the 2008 budget were held in December 2007.

[9]Laver and Benoit (2002) successfully used Wordscore to estimate TDs position in the 1991 confidence debate on the future of the Fianna Fáil–PD coalition government.

We assume that contributions in budget debates have the underlying dimension of being either *pro* or *contra* the current government. Our interpretation from reading the speeches is that, apart from the budget speech itself, all other contributions to a large extent either attack or defend the incumbent government. We therefore can use contributions during the budget debate as an indicator for how much a speaker is supporting or opposing the current government, here consisting of Fianna Fáil and the Green Party. As our reference texts we therefore chose the speeches by Bertie Ahern (Taoiseach) and Enda Kenny (FG party leader). The former should obviously be strongly supportive of the government while the latter, as party leader of the largest opposition party, should strongly oppose it. Figure 7 shows estimated positions for all speakers grouped by party affiliation.

[∗∗∗∗ FIGURE 7 ABOUT HERE ∗∗∗∗]

The estimated positions are clustered into two groups, one representing the government and one the opposition. Within the government cluster, Deputy Batt O'Keeffe (Minister of State at the Department of Environment, Heritage and Local Government) is estimated to be the most supportive speaker for the government, while Deputy Pat Carey (Minister of State at the Department of Community, Rural and Gaeltacht Affairs) and Deputy Sean Ardagh are estimated to be relatively closer to the opposition. Deputy John Gormley, leader of the Green party and Minister for the Environment, Heritage and Local Government in the FF-Green coalition, is estimated to be in the centre of the government cluster. Among all positions in the opposition cluster, the speech of Róisín Shortall is the closest to the government side, with Neville being the farthest out.

3.4 Ministers Policy Position in the 26th Government

The government cabinet in parliamentary democracies is at the core of political decision-making, yet it is difficult to model intra-cabinet bargaining as the preferences of most cabinet members are unknown. Cabinet decisions are usually made behind closed doors and the doctrine of joint cabinet responsibility prevents ministers from publicly opposing decisions, even if they disagree with them. Using ministers' speeches and their responses during question times offer a unique opportunity to infer their preferences on policy dimensions of interest. In our final application we estimate policy positions for all cabinet members in the 26th government. The dimension on which positions are estimated represents pro- versus contra-government spending. We show that estimated positions are highly correlated with departments' actual spending, which means that estimated positions are not only meaningful but also can be used to predict actual policy-making.

The 26th government was formed as a coalition between Fianna Fáil and the Progressive Democrats after the election for the 29th Dáil in 2002. The cabinet was reshuffled on 29 September 2004 and we only include ministers' speeches until that date. Table 2 lists all cabinet members (and their portfolios) included in our analysis.

[$****$ TABLE 2 ABOUT HERE $****$]

To estimate ministers' policy positions, we retrieve the complete record of each minister's contribution in parliament from the first meeting on 6 June 2002 until the date of the reshuffle. On average, each minister made 3,643 contributions with an average number of 587,077 words. Table 3 provides summary statistics for all ministers, sorted by total word count.

[$****$ TABLE 3 ABOUT HERE $****$]

We again use Wordscore (Benoit and Laver, 2003; Laver, Benoit and Garry, 2003) to estimate positions as it allows us to define the underlying policy dimension by choosing appropriate reference texts. We estimate positions on a social-economic left-right dimension that reflects pro- versus contra-government spending. We therefore use contributions by Mary Coughlan (Minister for Social and Family Affairs) and Charlie McCreevy (Minister for Finance) as reference texts, assuming that the former is more in favor of spending than the latter.

Figure 8 shows the results of estimation grouped by the two parties.

[∗∗∗∗ FIGURE 8 ABOUT HERE ∗∗∗∗]

As expected, we find that the two PD members, Mary Harney and Michael McDowell, are at the right side of the dimension. We estimate the most left-wing members to be Éamon Ó Cuív (Minister for Community, Rural and Gaeltacht Affairs), Noel Dempsey (Minister for Education and Science) and Micheál Martin (Minister for Health and Children). The most right-wing members are John O'Donoghue (Minister for Arts, Sport and Tourism) and Charlie McCreevy (whose contributions we used as right-wing reference text), Michael Smith (Minister for Defence).

How valid are these estimated positions? In order to have substantive meaning, our estimates should be able to predict political decisions on the same policy dimension. We therefore use ministers' estimated positions to predict their departmental spending level (see Giannetti and Laver, 2005, for a similar analysis with data from Italy). Our dependent variable is each department's spending as share of the total budget in 2004. Our independent variable is the vector of estimated policy positions. We conjecture that more left-wing ministers should have higher spending levels than right-wing ministers, which we test by

estimating

$$\widehat{\text{spending}} = \hat{\beta}_0 + \hat{\beta}_1 \text{policy position} \tag{1}$$

via ordinary least-square regression.

Figures 9 and 10 show the two variables plotted against each other together with the estimated regression line from equation 1. In Figure 9 we include all cabinet members. In Figure 10 we exclude non-spending departments with small budgets, such as the office of the Taoiseach or the Department of Foreign Affairs.[10]

[∗∗∗∗ FIGURE 9 ABOUT HERE ∗∗∗∗]

[∗∗∗∗ FIGURE 10 ABOUT HERE ∗∗∗∗]

Figure 9 reveals that there is a negative, albeit weak relationship between estimated positions and spending, with more left-wing cabinet members having higher spending levels than right-wing members. The correlation between the two variables is -0.53 ($p = 0.0523$) but which is not significant at the 0.05 level. However, if we only take members from high-spending departments into account (Figure 10) we find an almost perfect linear relationship between the two variables with a correlation coefficient of -0.95 ($p = 0.0002$). This result demonstrates that our estimated positions are indeed meaningful and can be used to predict behavior on political decisions on the same dimension.[11]

[10]The eight high-spending departments we include are, in decreasing order of budget share, the Department of Health and Children, Department of Education and Science, Department of Social and Family Affairs, Department of the Environment and Local Government, Department of Transport, Department of Enterprise, Trade and Employment, Department of Defence, and the Department of Arts, Sport and Tourism. These eight departments together account for more than 95 per cent of the total budget in 2004.

[11]These results also open up an intriguing question about the endogeneity of observable policy preferences of ministers. Do higher spending portfolios receive more left-leaning ministers or do ministers adapt their policy preferences after appointment and literally grow into the job? This and related questions are outside the scope of this paper and will be pursued by authors in future research.

4 Conclusion

Policy preferences of individual politicians (ministers or TDs in general), are inherently unobservable. However, we have abundant data on speeches made by political actors. The latest developments in automated text analysis techniques allow us to estimate policy positions of individual actors from these speeches.

In relation to Irish political actors such estimation has been hindered by the structure of the available data. While all speeches made in Dáil Éireann are dutifully recorded, the architecture of the data set, where digitised versions of speeches are stored, makes it impossible to apply any of the existing text analysis software. Speeches are currently stored by the Dáil Éireann in more than half a million separate HTML files with entries that are not related to each other.

In this paper we present a new database of speeches that was created with the purpose of allowing the estimation of policy preferences of individual politicians. For that reason we created a relational database where speeches are related to the members database and structured in terms of dates, topics of debates, and names of speakers, their constituency and party affiliation. This gives the necessary flexibility to use available text scaling methods in order to estimate policy positions of actors.

We also present several examples for which this data can be used. We show how to estimate policy positions of all Irish Ministers for Finance, and highlight how this can lead to interesting research question in estimating the determinants of their positions. We show that for some ministers the position can be explained by the economic performance of Ireland, while preferences of other ministers seem to be idiosyncratic. In another example we estimate positions of individual TDs in a budget debate, followed by the estimation of policy

positions of cabinet members of the 26th Government.

With the introduction of our database, we aim to make text analysis an easy and accessible tool for social scientist doing empirical research on policy-making in Ireland that requires estimation of policy preferences of political actors.

References

Baayen, R.H. 2001. *Word frequency distributions*. Vol. 18 of *Text, Speech and Language Technology* Springer.

Benoit, K. and M. Laver. 2003. "Estimating Irish party positions using computer wordscoring: The 2002 elections." *Irish Political Studies* 17(2):97–107.

Bybee, J.L. 2001. *Phonology and language Use*. Cambridge University Press Cambridge.

Charbonneau, Etienne. 2009. Talking Like a Tax Collector or a Social Guardian? The Use of Administrative Discourse by US State Lottery Agencies. In *Do They Walk Like They Talk?: Speech and Action in Policy Processes*, ed. Louis Imbeau. Springer pp. 223–241.

Christen, Peter. 2006. A Comparison of Personal Name Matching: Techniques and Practical Issues. Joint computer science technical report series Department of Computer Science, The Australian National University.

Clinton, J. D., S. Jackman and D. Rivers. 2004. "The Statistical Analysis of Roll Call Voting: A Unified Approach." *American Political Science Review* 98(2):355–370.

Condon, Stephanie. 2009. "Word Cloud Of Obama And Cheney Speeches." *CBS News, May 21* .

Day, Kate. 2009. "Barack Obama's inauguration speech as a tag cloud." *Telegraph, January 20* .

Galli, Emma, Veronica Grembi and Fabio Padovano. 2009. Whould You Trust an Italian Politician? Evidence from Italian Regional Politics. In *Do They Walk Like They Talk?: Speech and Action in Policy Processes*, ed. Louis Imbeau. Springer pp. 109–131.

Gavin, Patrick. 2009. "Word Cloud: Obama's Speech." *Politico.com, September 9* .

Giannetti, Daniela and Michael Laver. 2005. "Policy positions and jobs in the government." *European Journal of Political Research* 44(1):91–120.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18(1):1–35.

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

Imbeau, Louis. 2009. Dissonance in Fiscal Policy: A Power Approach. In *Do They Walk Like They Talk?: Speech and Action in Policy Processes*, ed. Louis Imbeau. Springer pp. 167–185.

Kluver, H. 2009. "Measuring interest group influence using quantitative text analysis." *European Union Politics* 10(4):535–549.

Laver, M., K. Benoit and J. Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97(2):311–331.

Laver, Michael and Kenneth Benoit. 2002. "Locating TDs in policy spaces: Wordscoring Dáil speeches." *Irish Political Studies* 17(1):59–73.

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

Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. "Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16(4):372–403.

Proksch, Sven-Oliver and Jonathan B. Slapin. 2009*a*. "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. 2009*b*. "Position taking in European Parliament speeches." *British Journal of Political Science* forthcoming.

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

Slapin, J. B. and S.-O. Proksch. 2008. "A scaling model for estimating time-series party positions from texts." *American Journal of Political Science* 52(3):705–722.

Williams, Marc. 2008. "Analysing Cameron's 400,000 words." *BBC Political Research Unit, October 1* .

Yao, Laura. 2008. "We've Looked at Clouds From Both Sides Now." *The Washington Post, August 3* .

Yu, Bei, Stefan Kaufmann and Daniel Diermeier. 2008. "Classifying party affiliation from political speech." *Journal of Information Technology & Politics* 5(1):33–48.

Figure 1: *Word clouds of all Finance Ministers' budget speeches, 1922–2008*

Figure 1: *Word clouds of all budget speeches made by Ministers for Finance, 1922–2008 (cont'd).*

Figure 1: *Word clouds of all budget speeches made by Ministers for Finance, 1922–2008 (cont'd).*

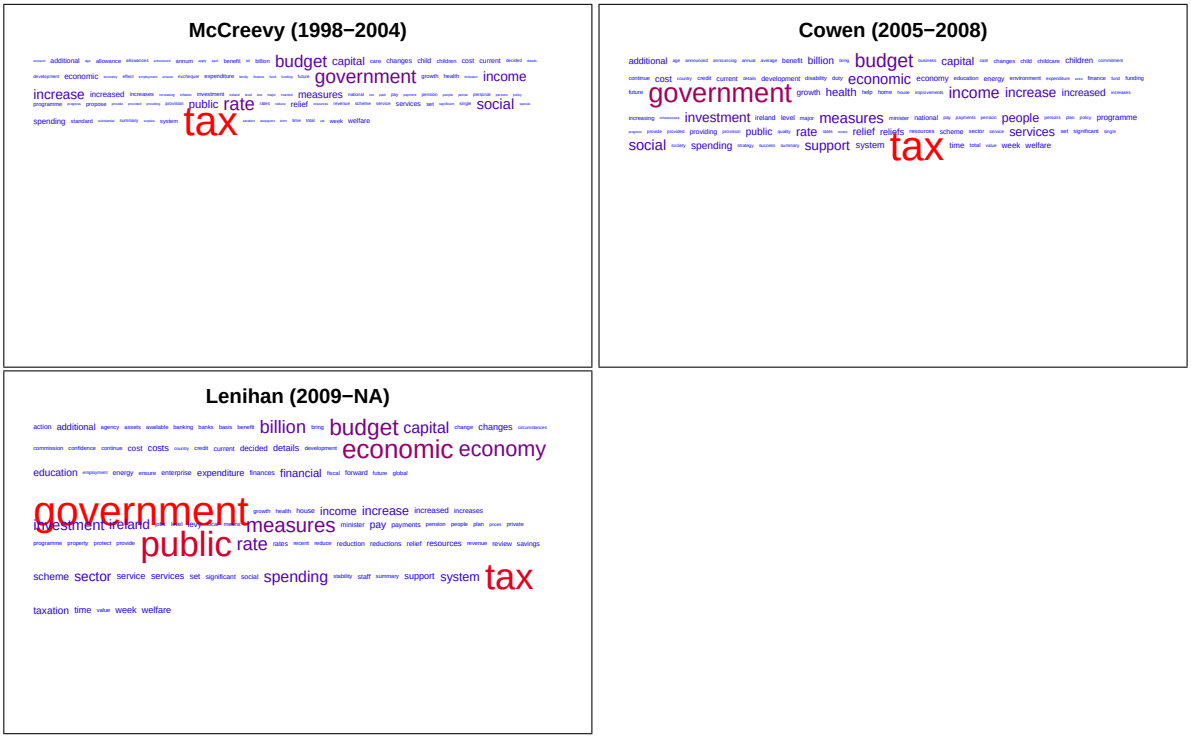

Figure 1: *Word clouds of all budget speeches made by Ministers for Finance, 1922–2008 (cont'd).*

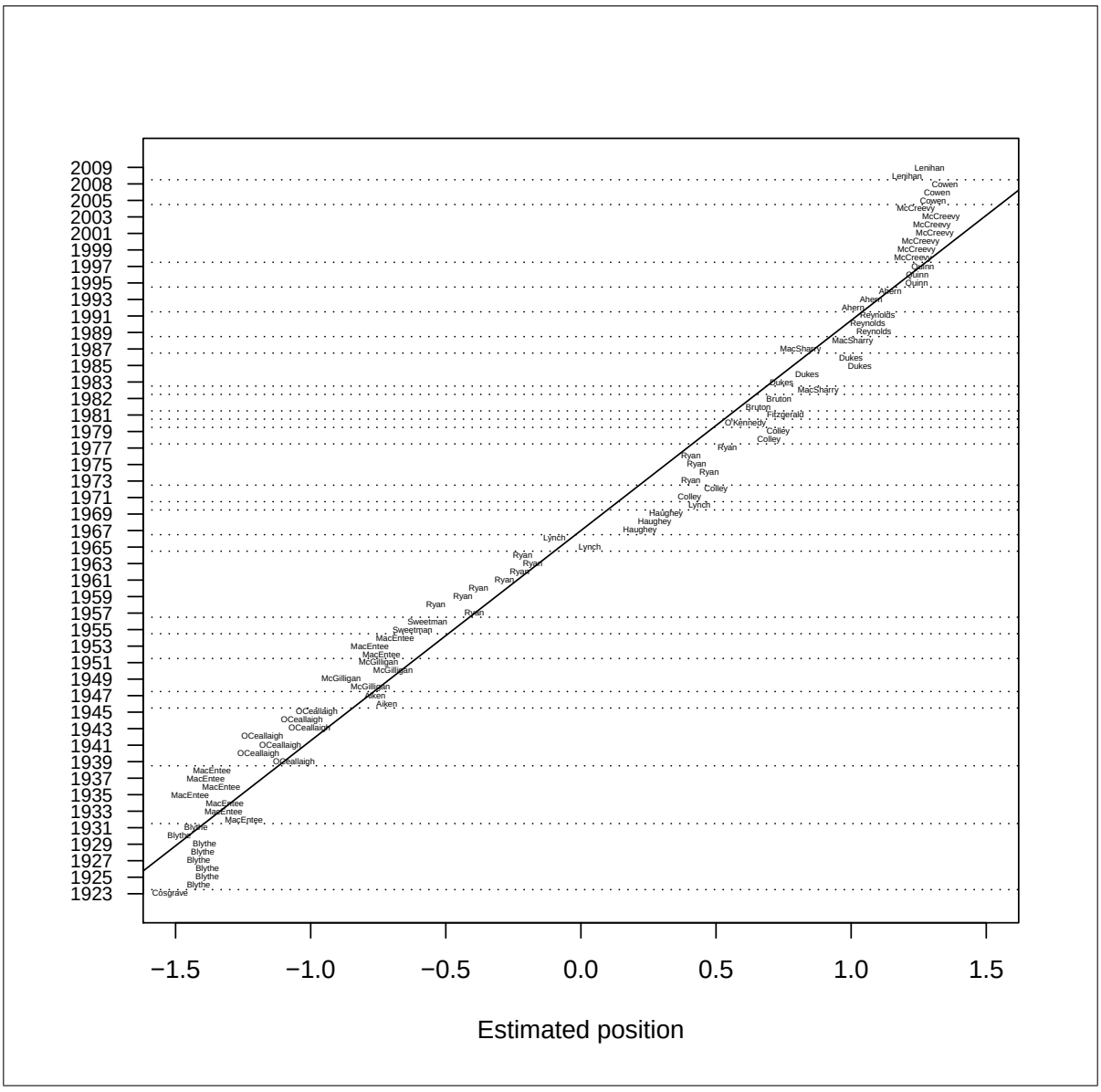

Figure 2: *Finance ministers' policy positions as estimated from all budget speeches (1922–2009) with an overlaid linear regression line.*

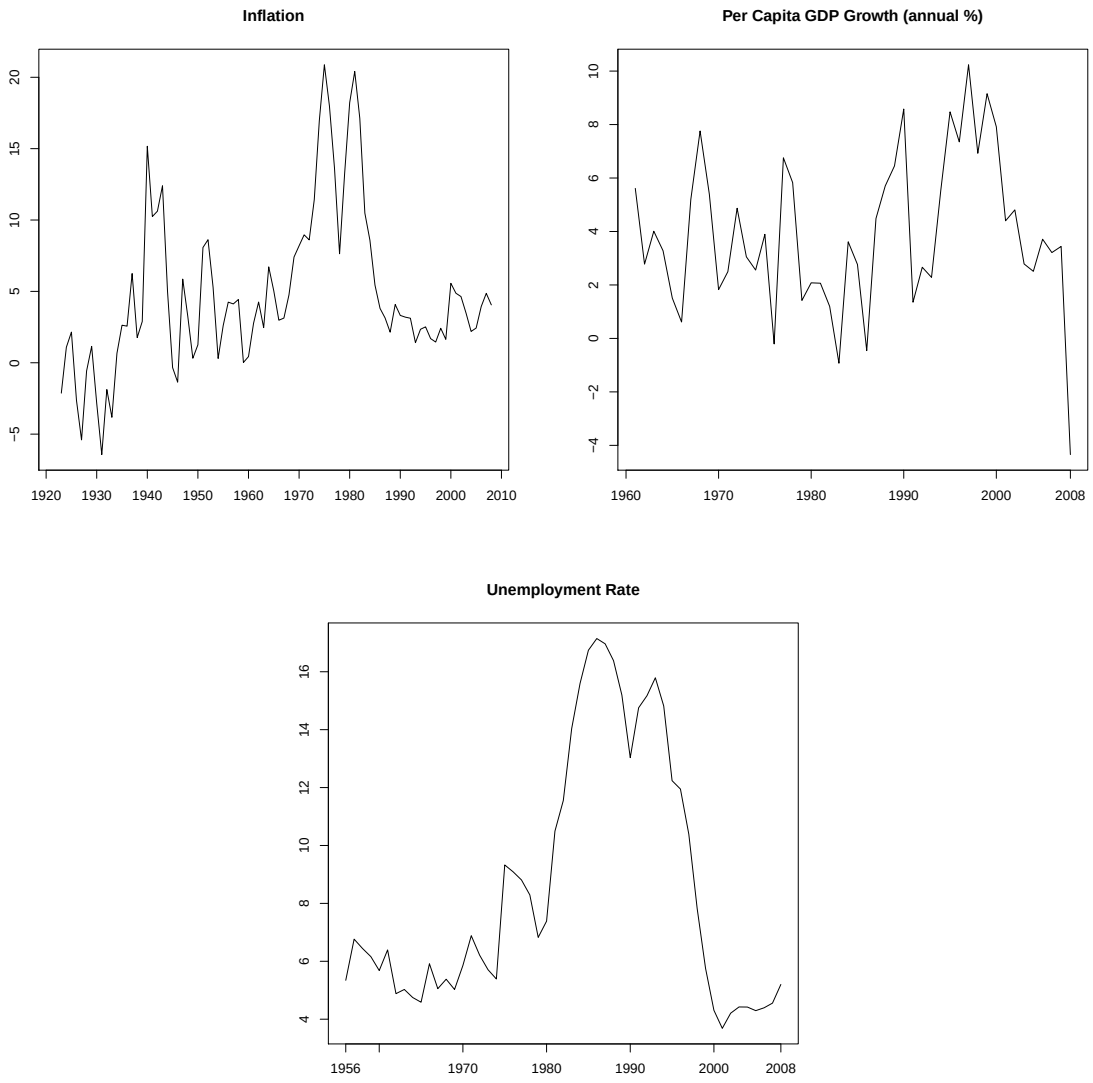

Figure 3: *The Irish economy over time: Inflation (1923–2008), Per Capita GDP growth (annual %; 1961–2008) and unemployment rate (1956–2008).*

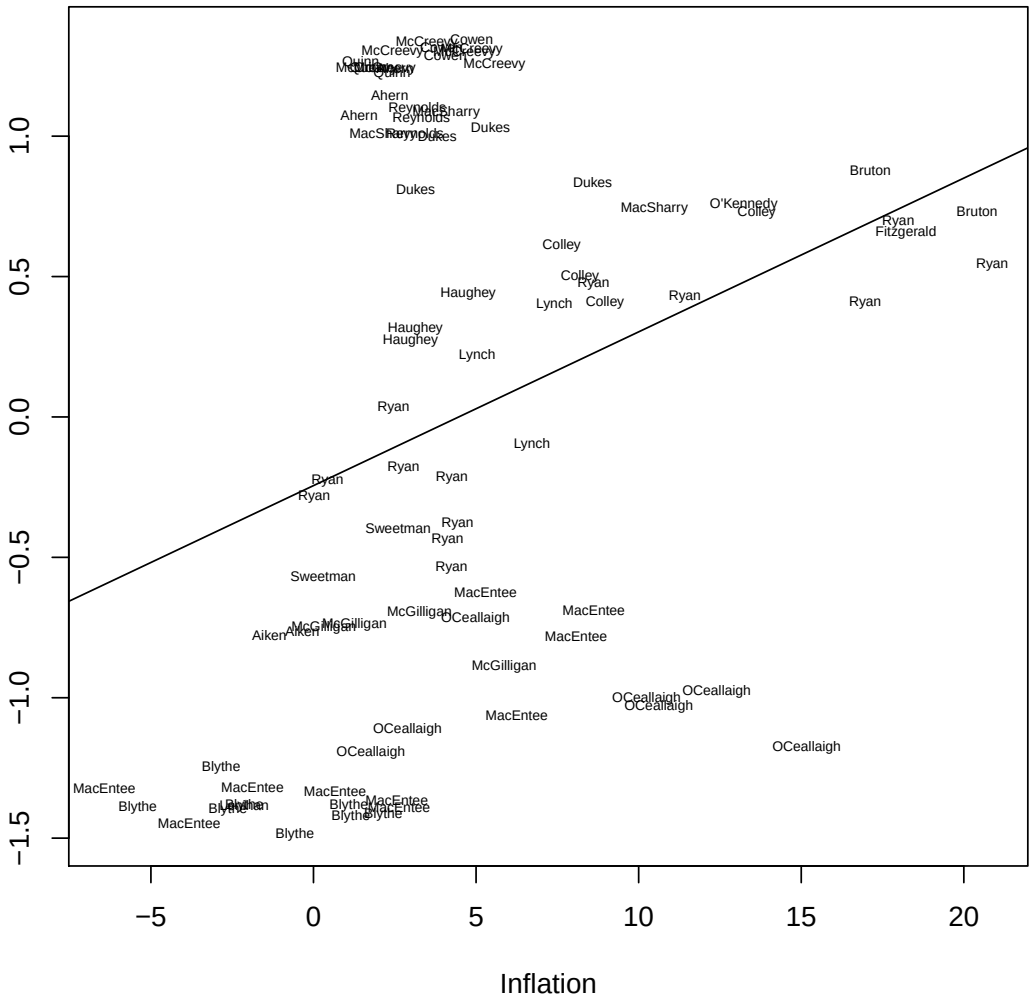

Figure 4: *Estimated finance ministers' positions against inflation (1923–2008) with an overlaid linear regression line.*

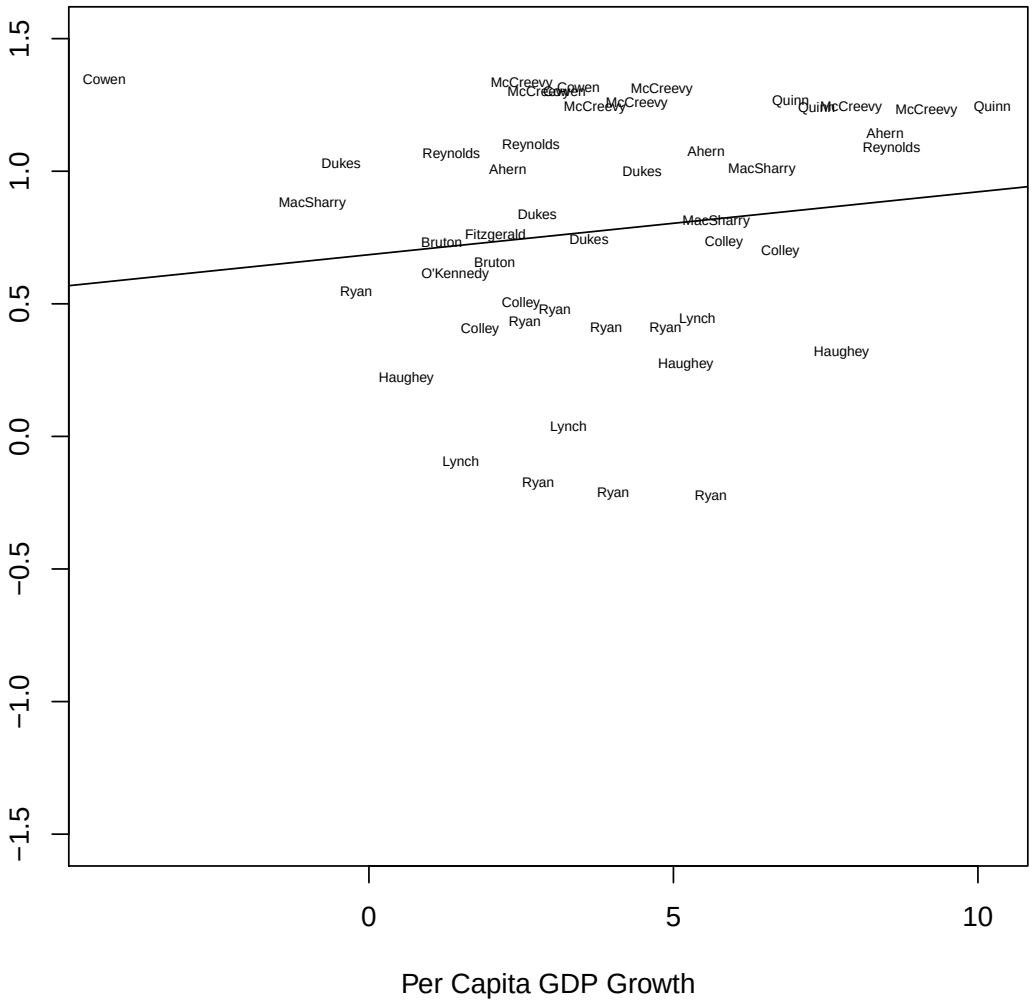

Figure 5: *Estimated finance ministers' positions against GDP growth (annual %; 1961–2008) with an overlaid linear regression line.*

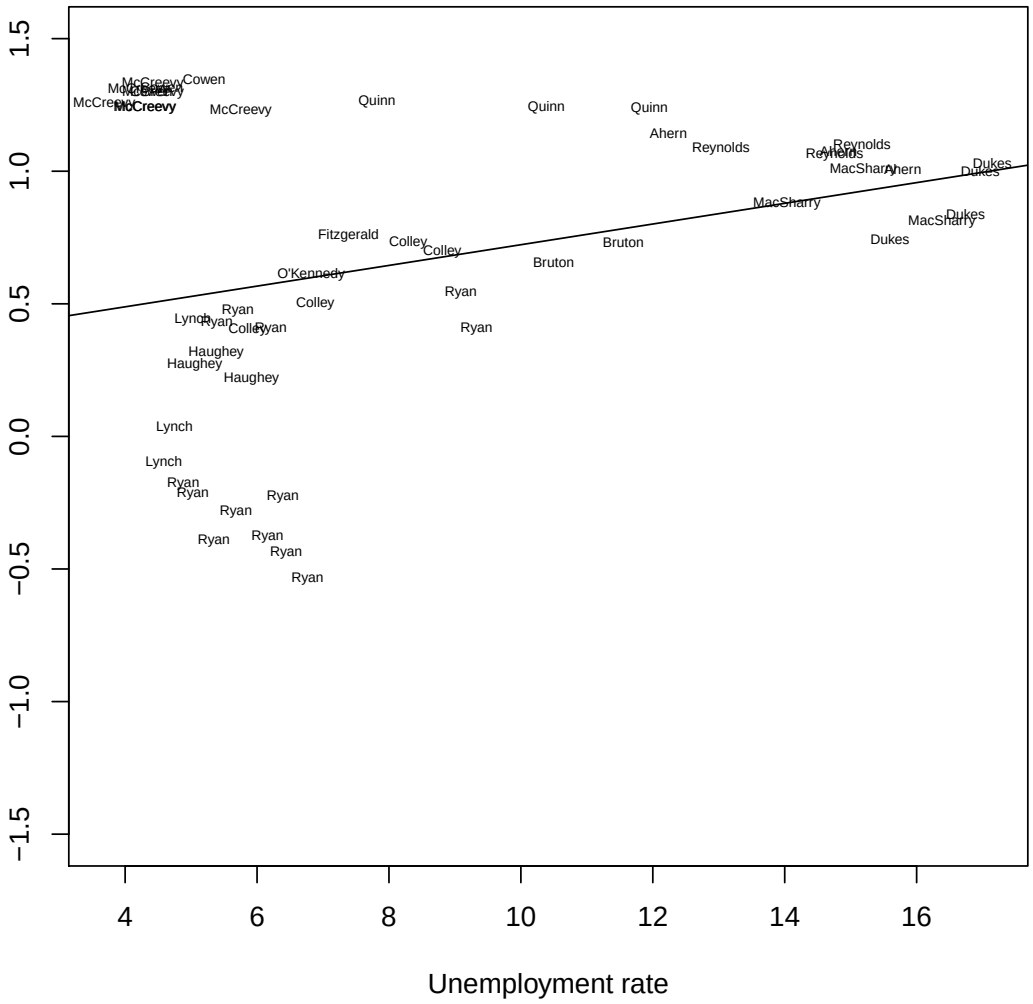

Figure 6: *Estimated finance ministers' positions against unemployment rate (1956–2008) with an overlaid linear regression line.*

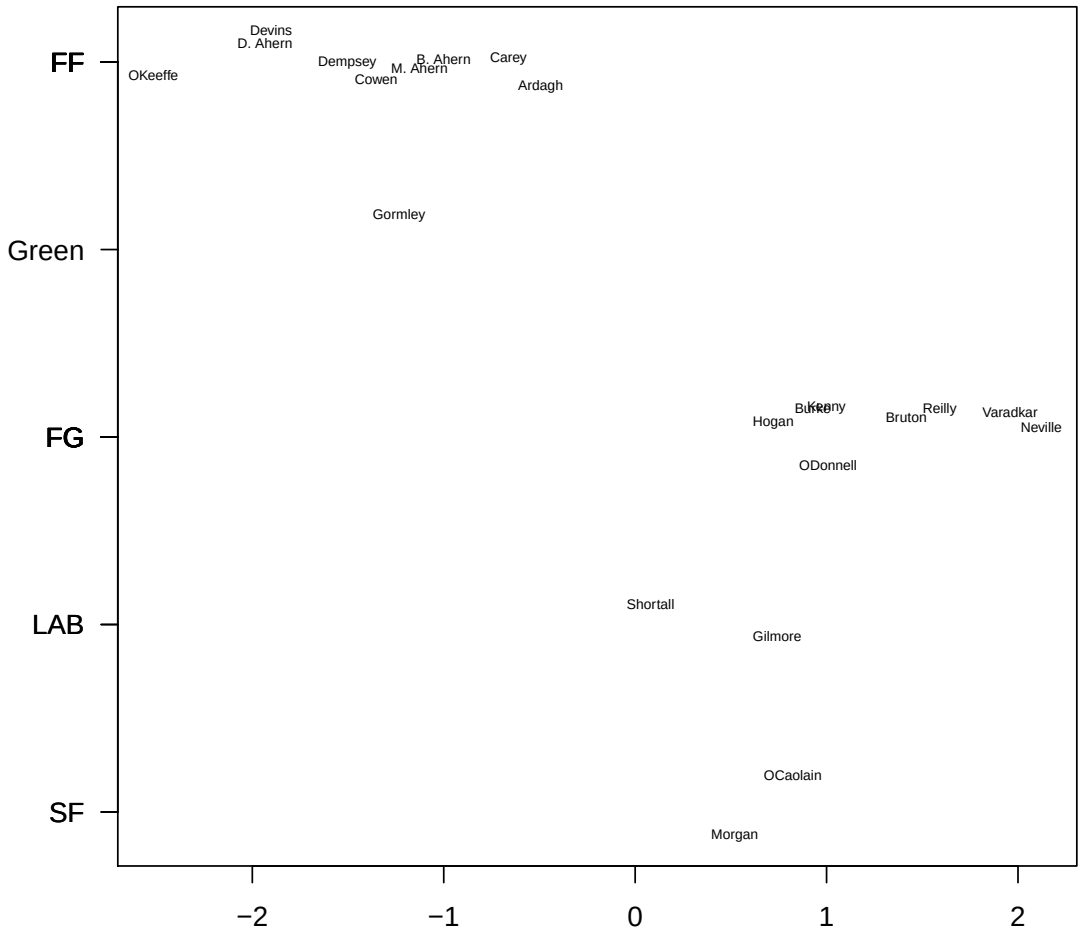

Figure 7: *Estimated positions of all speakers in the 2008 budget debate. Estimated dimension represents pro- versus anti-government positions. Scaling of x-axis is arbitrarily. Speeches of Bertie Ahern (FF, Taoiseach) and Enda Kenny (FG party leader) were used as reference texts for being respectively pro- or anti-government. Observations are jittered along the y-axis to prevent names from overlapping.*

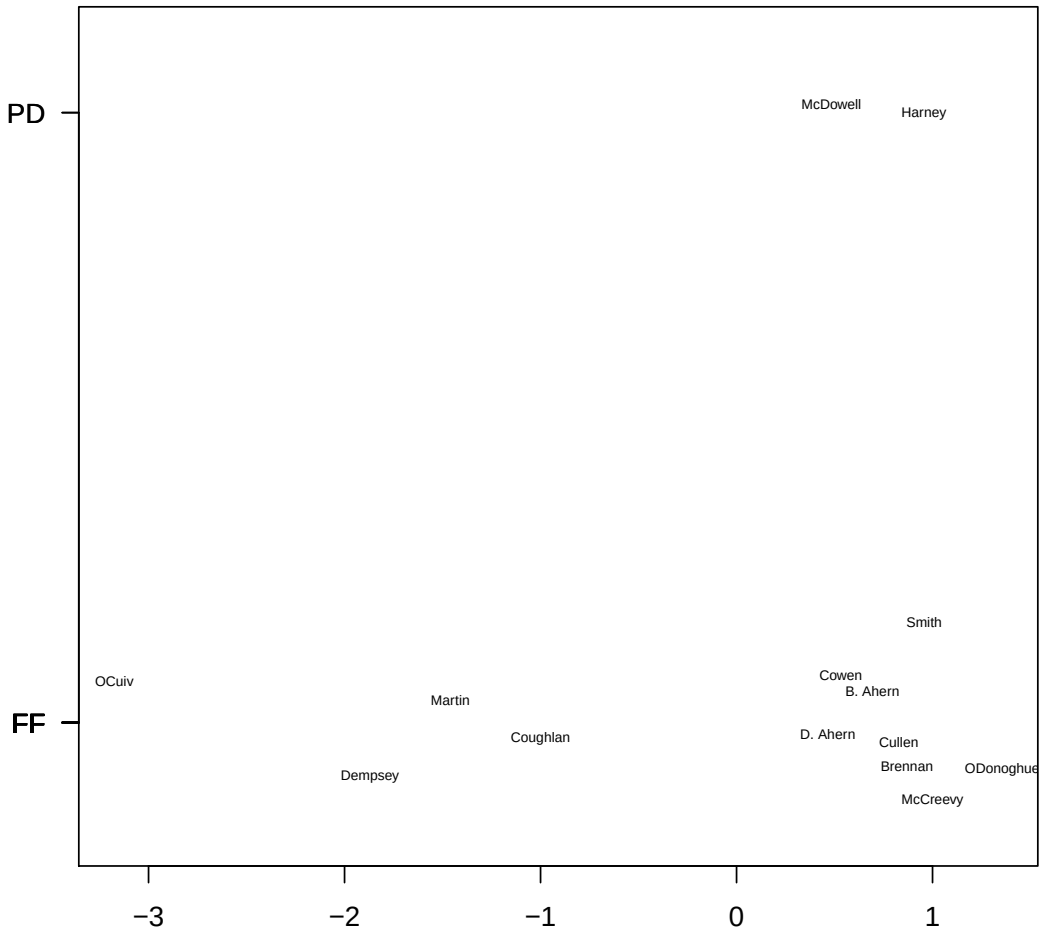

Figure 8: *Estimated positions for all cabinet members in the 26th government (29th Dáil) using Wordscore. Positions are jittered along the y-axis. Estimation is based on each minister's contribution in Dáil Éireann before the cabinet reshuffle on 29 September 2004. Speeches by Mary Coughlan (Minister for Social and Family Affairs) and Charlie McCreevy (Minister for Finance) are used as left and right reference texts, respectively*

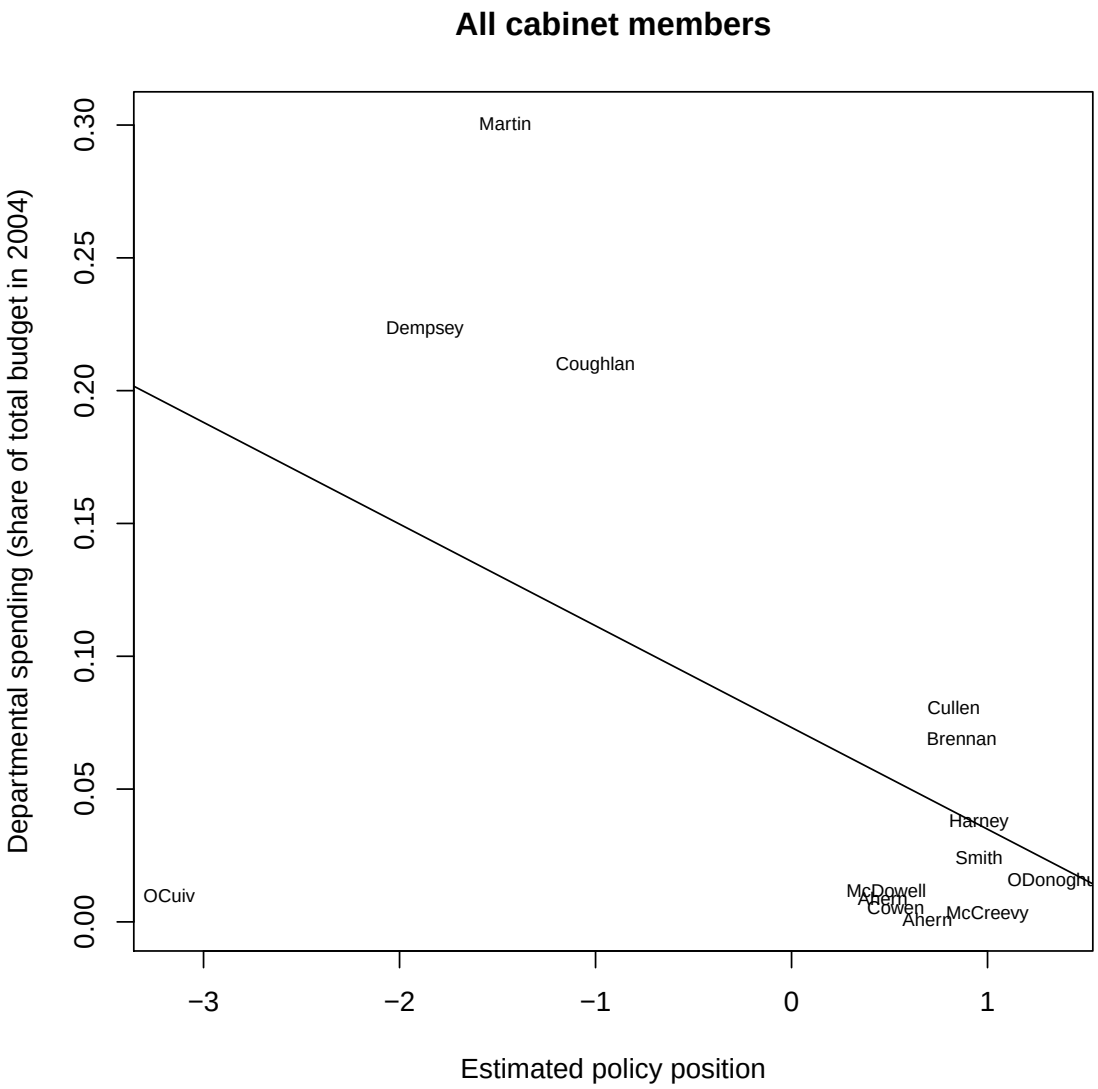

Figure 9: *Cabinet ministers' policy position plotted departmental spending as share of total government budget in 2004.*

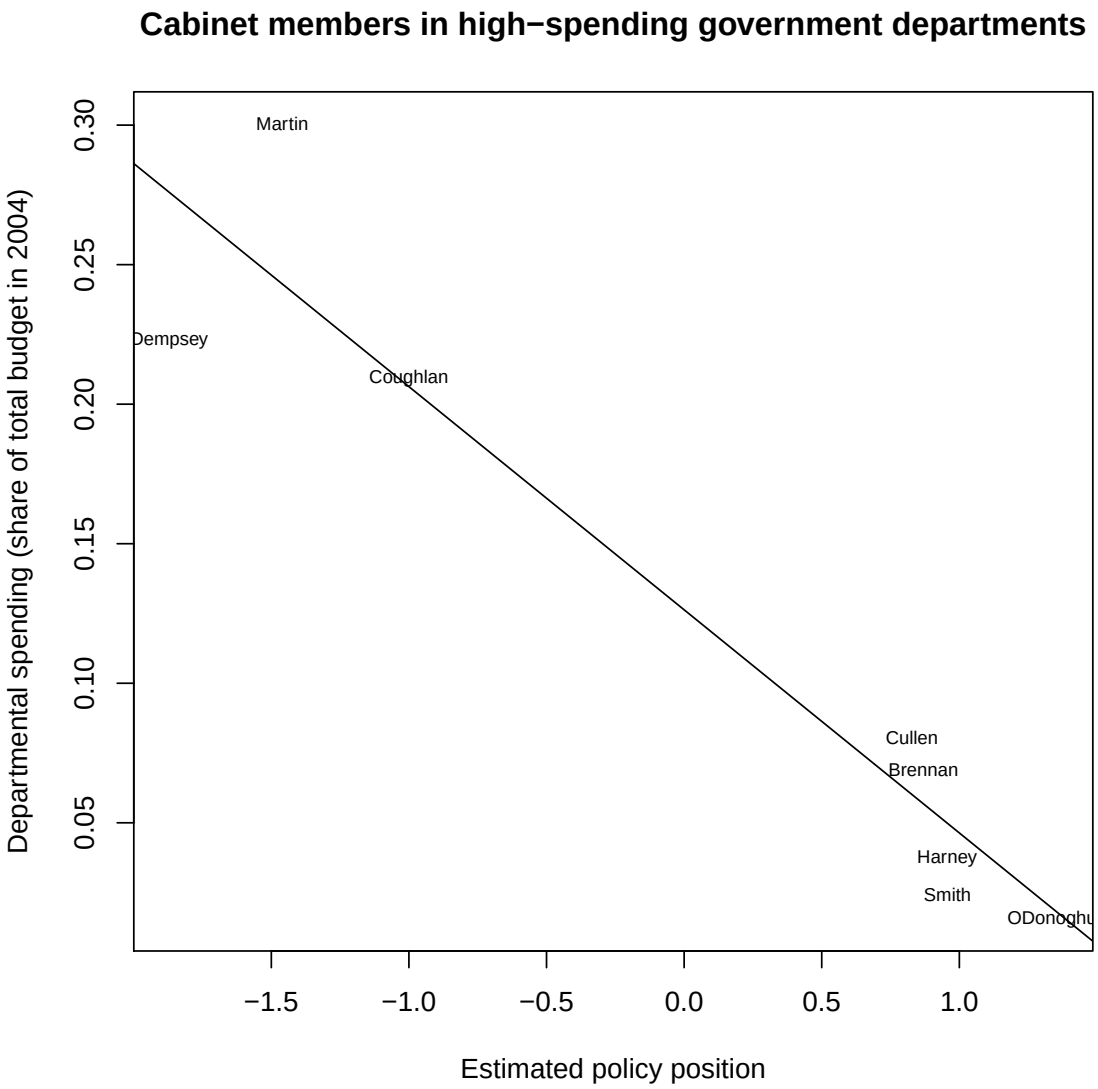

Figure 10: *Cabinet ministers' policy position plotted against departmental spending as share of total government budget in 2004. Low-spending departments such as the office of the Taoiseach or the Department of Foreign Affairs have been excluded. The remaining eight departments together account for more than 95 per cent of the total budget in 2004.*

Table 1: *Speakers in the 2008 budget debate.*

Name	Party	Government party	Length of speech in number of words
Ahern, Bertie[1]	FF	Yes	3,959
Ahern, Dermot	FF	Yes	2,700
Ahern, Michael	FF	Yes	1,190
Ardagh, Sean	FF	Yes	1,015
Carey, Pat	FF	Yes	942
Cowen, Brian[2]	FF	Yes	8,733
Dempsey, Noel	FF	Yes	1,438
Devins, Jimmy	FF	Yes	1,090
O'Keeffe, Batt	FF	Yes	715
Gormley, John	Green	Yes	4,306
Bruton, Richard	FG	No	10,817
Burke, Ulick	FG	No	714
Hogan, Phil	FG	No	1,438
Kenny, Enda[3]	FG	No	3,924
Neville, Dan	FG	No	1,210
O'Donnell, Kieran	FG	No	1,182
Reilly, James	FG	No	1,683
Varadkar Leo	FG	No	1,876
Gilmore, Eamon	Labour	No	5,141
Shortall, Roisin	Labour	No	2,662
Morgan, Arthur	SF	No	6,158
O'Caolain, Caoimhghin	SF	No	1,438

Notes: 1–Taoiseach, 2–Minister for Finance, 3–FG Party Leader. The budget debate for the 2008 budget was held in December 2007.

Table 2: Members of the 26th government (29th Dáil), 6 June 2002–29 September 2004

Name	**Party**	**Office**
Bertie Ahern	FF	Taoiseach
Mary Harney	PD	Tánaiste and Minister for Enterprise, Trade and Employment
Michael Smith	FF	Minister for Defence
Joe Walsh	FF	Minister for Agriculture and Food
Charlie McCreevy	FF	Minister for Finance
Brian Cowen	FF	Minister for Foreign Affairs
Noel Dempsey	FF	Minister for Education and Science
Dermot Ahern	FF	Minister for Communications, Marine and Natural Resources
John O'Donoghue	FF	Minister for Arts, Sport and Tourism
Micheál Martin	FF	Minister for Health and Children
Séamus Brennan	FF	Minister for Transport
Michael McDowell	PD	Minister for Justice, Equality and Law Reform
Martin Cullen	FF	Minster for the Environment and Local Government
Éamon Ó Cuív	FF	Minister for Community, Rural and Gaeltacht Affairs
Mary Coughlan	FF	Minister for Social and Family Affairs

Source: Houses of the Oireachtas (http://www.oireachtas.ie/viewdoc.asp?DocID=2935).

Table 3: Summary statistics for ministers' contributions in the 26th government, 6 June 2002 – 29 September 2004, sorted by total word count

Name	**Party**	**Number of contributions**	**Total word count**
Noel Dempsey	FF	8,066	1,273,835
Michael McDowell	PD	6,290	1,038,527
Bertie Ahern	FF	6,505	790,964
Dermot Ahern	FF	3,047	755,471
Charlie McCreevy	FF	3,249	657,010
Brian Cowen	FF	2,444	652,062
Martin Cullen	FF	5,826	574,464
Séamus Brennan	FF	3,324	513,938
Mary Coughlan	FF	2,627	503,413
Mary Harney	PD	3,357	418,745
Michael Smith	FF	2,464	330,575
Éamon Ó Cuív	FF	1,459	286,194
John O'Donoghue	FF	1,553	282,154
Micheál Martin	FF	789	141,721

Note: Only speeches before the cabinet reshuffle on 29 September 2004 are included.

Institute for International Integration Studies
The Sutherland Centre, Trinity College Dublin, Dublin 2, Ireland