Lost in Translation? Predicting Party Group Affiliation from European Parliament Debates*

Bjorn Hoyland[1], and Jean-François Godbout[2]

September 19, 2008

[1] Department of Political Science, University of Oslo
[2] Department of Political Science, Simon Fraser University

ABSTRACT

Ten countries joined the European Union in 2004. This offered a rare opportunity for the existing party groups to substantively increase their share of the seats in the European Parliament by recruiting national party delegations from the new member states. As most of new member states have a relative short history of competitive multiparty system, there were weaker ties between parties in new and old member states when compared to previous rounds of enlargement. Since the allocation of office spoils in the EP is fairly proportional amongst party groups ,(Mamadouh and Raunio, 2003) it was assumed that national parties from the new member states – less ideologically committed to any of the belief systems held by the traditional Western European party families– would shift the allocation of some offices in the EP by opting to join certain party group who controlled a larger share of office spoils. In this paper we provide a novel approach to evaluate whether the party group choice of the national parties from the new member states may have been affected by such considerations. By employing a *Support Vector Machine*, a common classification tool from computational linguistics, we evaluate to what extent it is possible to correctly predict the party group affiliation of participants in European Parliament debates on the basis of the content of their speeches. Our results show that the differences in belief systems between party groups as expressed by Members of the European Parliament (MEPs) plenary session speeches in the 5^{th} (1999-2004) European Parliament are helpful in predicting the party group affiliation of MEPs from older member states in the 6^{th} European Parliament. However, these difference are not helpful when trying to predict the party group affiliation of MEPs from the new member states. This suggests that the national parties of these new member states opted to join the more established party groups for other reasons than similarity in belief systems.

*We would like to thank Indraneel Sircar for assistance in collecting the speech data and Bei Wu for help with the initial analysis of the data. All errors remain our own. Bjorn Hoyland would like to acknowledge the generous financial support provided by his department.

Legislatures aim to produce two essential tasks in democratic societies; legislate and provide a forum for the debate over policy alternatives. While the literature on the European Parliament (EP) suggests that it is a powerful legislature, most scholars agree that EP plenary sessions do not provide a forum for European public debates (Eriksen and Fossum, 2000). This can be explained by several factors. First, Members of the European Parliament (MEPs) are allowed to speak in any of the official languages of the Union. The speech is immediately translated and communicated to the rest of the chamber and to the audience with headphones. This small translation delay means that there is little room for spontaneity, unlike in other legislatures like the British Westminster Parliament or the U.S. Congress. Second, voting does not normally occur on the same day as the debate. And third, the debates are usually poorly attended. This probably explains why there is very little media attention focusing on the debate of the EU Parliament. There is little reason to expect that party groups exercise strong control over the content of individual speeches in the debates of the plenary as it is neither likely to change the outcome of the vote nor to be reported widely in the press. The debates do however relate directly to the policy issues facing the legislators. Furthermore, speeches allow for far more nuances than voting decisions. These debate could thus provide available source of information about politics in the European Parliament.

Text is becoming an increasingly common source of data in political science. There has been several innovate projects proposing new methods for extracting actors ideal points on the basis of the text (Laver and Garry, 2000; Laver, Benoit and Garry, 2003; Monroe and Maeda, 2004; Slapin and Proksch, 2008). Our project is less ambitious, we are simply interested in correctly classifying the ideological party group affiliation of members in the European Parliament. Instead of developing custom software for the classification of political science text (Hopkins and King, 2007), we borrow an existing classification tool from computational linguistics, *Support Vector Machine* (Baayen, 2008).

In this paper, we utilize debates to evaluate to what extent national parties from the new member states joined party groups with whom they share similar belief systems, as expressed through choice of words in plenary debates. We use a common classification algorithm, *Support Vector Machine*, to predict party group affiliation of EP plenary speakers in the first three years of the 6^{th} European Parliament (July 2004 - July 2007) on the basis of the party group affiliations of the legislators in all plenary debates in the 5^{th} European Parliament. Our criteria is straightforward: if the *Support Vector Machine* does a better job at correctly classifying MEPs from the old member states than MEPs from the new member states, then we have support for the claim that ideological principles were not the key factors behind the Eastern European parties' affiliation choices.

The paper is structured as follows. The first section reviews the literature on the role of ideology in the European Parliament. We show that ideology has been measured with a multitude of methods, like voting behavior, MEPs surveys, expert opinion studies, and recently, also plenary debates. The second section surveys the development of computational methods to analyze political texts. The third section presents the data and the practical details for preparing the analysis. The results are presented in the final section. The key finding is that the *Support Vector Machine* help us to correctly predict the party group affiliation of speakers in the 6^{th} European Parliament. In particular, it does a significantly better job at predicting the party group affiliation of MEPs from Western Europe than MEPs from Eastern Europe. This is even the case when MEPs present in the 5^{th} Parliament were excluded from the analysis. Our findings suggest that the parties from the new member states were all but guided by ideology or similarity in belief system with parties from the old member states in their choice of party group affiliation.

IDEOLOGY IN THE EUROPEAN PARLIAMENT

The role of ideology inside the European Parliament is a hotly debated topic. Although the European Parliament has organized itself into transnational political party groups ever since the birth of the institutions, scholars have questioned the extent to which the European Parliament offers a forum for European wide debate (Eriksen and Fossum, 2002). Practitioners have also called for a politicization of the activities of the European Parliament in order to highlight the ideological, rather than the technical, dimension of EU legislation (Clegg and van Hulten, 2003). Until recently, quantitative EP scholars have not used the debates as a source for obtaining information about the ideology of MEPs. Instead, information about their ideology have been obtained in a more indirect manner.

Hix (2002) uses self-placement on left-right and pro/anti integration scales as indicators of ideology. NOMINATE estimates obtained from roll-call votes are also used as a measure of "revealed" preferences (for a critique, see Carrubba et al., 2006). The general consensus seems to be that the left-right ideology continuum is capable of explaining a substantive proportion of voting behavior on roll call votes in the European Parliament.

A central debate in the literature on legislative behavior in the European Parliament is to what extent transnational party groups are able to discipline the revealed behavior of their members. This is complicated by the fact that these transnational party groups are composed of national party delegations. It is extremely rare that individual members defect from their national party delegation to vote with their own party group when there is intra-party conflict (Hix, 2004). In addition, in most roll call votes, we frequently find that all MEPs of a national party delegation simply vote with their own cross-national party group.

As a consequence, scholars have found that party groups in the EP are highly cohesive (Hix, Noury and Roland, 2005). What remains unclear is why newly admitted MEPs chose to join these more established parties groups in the first place. One potential explanation is related to rent seeking behavior: since older party groups formally control the office spoils within the EP, representatives from newly admitted states would join these parties to gain broader access to parliamentary committees for example. Another explanation is that national party delegations joined the more established party groups because they shared a similar ideology. Recent work by Hix and Noury (2006) has shown that voting behavior in the European Parliament has barely changed after the enlargement of the union in 2004. Using roll call votes from the first half of the 6^{th} European Parliament, the authors compared the voting behavior of new member delegations with members from the 5^{th} Parliament. This study concludes that transnational party lines remains the main predictor of roll call voting, and that voting along national lines is still a relatively rare event.

Although it may be surprising to some readers, it appears that party group cohesion in roll call votes does not suffer from decreased ideological cohesion as measured by exogenous variables from MEPs surveys and expert opinions (Hix, Noury and Roland, 2007). As was indicated earlier, this cohesion did not follow a notable decline after the Eastern and Southern enlargement in 2004, when ten new countries joined the union. This is surprising given the fact that most of the new member states do not have particular stable national party systems unlike their Western European counterparts. Comparisons of survey responses and roll call votes show less variation within party groups in the revealed voting behavior than in the survey. This is the case for MEPs from both new and old member states (Farrell et al., 2006; Gabel and Hix, 2007). This may be due to the fact that survey questions were fairly general and not specific enough to address the policy-issues facing the legislators. In addition, the survey questions may be interpreted differently by people with similar ideological orientation. Moreover, there is no clear link between the questions asked in the survey and the issues discussed in the European Parliament. Ideally, it would have been useful to contrast some measure of roll call votes with another activity undertaken by members of the European Parliament.

To sum up, we can see that the effect of the Eastern enlargement on politics in the European Union remains a contentious topic amongst EU scholars. Hix and Noury (2006) have analyzed voting behavior in the first 2 years after the Eastern enlargement. They find that parties form the new member states tend to vote with the transnational party group they are affiliation with. To the extent that voting reflects similar ideological world-views, or belief systems, party group affiliation seems to be a very good proxy. There are however reasons to be careful before equating voting similarities with similarities in belief systems. It may be the case that two individuals display similar voting record, not because they have similar preferences, ideology, or belief system, but because they just follow the same voting instructions in order to obtain their share of the office spoils.

COMPUTATIONAL ANALYSIS OF POLITICAL TEXTS

Political scientists have traditionally relied on labor intensive coding methodologies to study political texts, party platforms, government declarations, and campaign speeches (e.g. Budge et al., 2001; Klingemann et al., 2006). In the European Parliament, official records of debates are printed on a daily basis in large volume. The sheer size of these volumes renders any attempt to systematically code the content of legislative speeches highly impractical. However, recent developments in computer science have provided researchers with a series of tools that can be used to automatically process texts (for a recent review, see Cousins and McIntosh, 2005).

Computational approaches have been applied mainly in two areas of political science. First, it has been applied to study issues, topics and legislative agendas. The main focus of this application is to investigate how policy agendas change over time. Recent work have applied either fully or partially computerized approaches to process political texts. Some contributions rely on dictionary-based approaches (Laver and Garry, 2000). Here, topics are automatically classified using a manually coded dictionary. Other contributions rely on fully automated approaches (Purpura and Hillard, 2006; Quinn et al., 2006; Schonhardt-Bailey, 2006). An active application of automated text analysis is in the area of ideal point estimation. Laver, Benoit and Garry (2003) used a semi-automated approach in their analysis of party manifestos. First, they used human coding to classify a few parties as reference point on a left-right ideological scale. They then extracted word frequencies from their respective party manifestos. Other parties' ideological positions are then estimated based on how closely their party manifesto's word frequency distributions matched the reference distributions. An alternative approach was developed by Monroe and Maeda (2004). They use a statistical technique similar to Poole and Rosenthal (1997) in order to estimate "rhetorical ideal points". Their choice space, however, is a matrix of word counts rather than vote count for each legislator. Unfortunately, as they discuss in their paper, this new method faces various statistical and computational difficulties. Recently, Slapin and Proksch (2008) have proposed a multivariate fixed effect Poisson model, labeled *Wordfish* as an alternative approach to obtain party positions on the basis of text. Their approach has several desirable properties. First, *Wordfish* is capable of producing an ideal position of each actor on the estimated dimension, without the need for reference texts. Second, *Wordfish* also estimate the discrimination parameter of all words (or wordstems). This allow the researcher to identify the words that are best able to distinguish between actors. Unfortunately, their approach is computational intensive, in particular if the researcher would like to obtain the uncertainty associated with the location of actors and words. *Wordfish* has been applied to the debates in the plenary session of the 5^{th}

European Parliament. Using national party delegations as units, their results suggest that the debates in the European Parliament reflects divisions over European Integration rather than left-right politics (Proksch and Slapin, forthcoming).

Our approach is somewhat different. Rather than developing our own methodology for analyzing word count frequencies, we draw on recent development in computational linguistics. In this field, supervised classification tools are widely used and well developed. Supervised learning approaches use sets of text to train a classifier and then employ the trained classifying algorithm to classify a series of untrained text. The success of the classifying exercise is defined by how well it can classify a new set of documents. The best known example of supervised learning is perhaps email spam filter to identify junk mailings.

There are only a few applications of this approach in politics. Examples include studies by Purpura and Hillard (2006), who use supervised learning techniques to code speech topics in the U.S. Congress and Thomas, Pang and Lee (2006) who investigate (with modest success) whether speech classification of floor debate in the House of Representatives on a specific bill can be used as a predictor for subsequent agreement. Thomas et al. investigate whether we can predict a speaker's opinion toward a specific bill (support or opposition) based on his or her speeches. They used 2740 speech segments in 38 bill debates to train an SVM classifier and then use the classifier to predict the opinions of 860 speech segments in 10 other legislative debates. The plain SVM classifier achieves 66% prediction accuracy. Additional information of same-speaker links and inter-speaker agreement links helps improve the accuracy to 76%. Diermeier et al. (2006) train a Support Vector Machine to classify ideology in the US Senate. A detailed technical explanation of the technique can be found therein. In their paper, the authors analyze legislative speech records in the U.S. Senate from the 101st-108th Congresses. When predicting the conservative/liberal positions of Senators with previous legislative speech, the SVM classifier achieved a 94% level of accuracies. The authors conclude that the same ideological differences associated with roll-call based ideological measures in Congress are reflected in political speech.

Like Diermeier et al. (2006), we use the Support Vector Machine (SVM) supervised learning algorithm to classify MEPs' party affiliation by the content of their speech. SVM is a supervised learning method based on the the Structural Risk Minimization principle from statistical learning theory (Vapnik, 1998, 1999). The data, consisting of documents of all legislative speech delivered by MEPs, corresponds to vectors in an n-dimensional space, where each dimension represents a relevant feature in the classification task.

As the European Parliament is a multi-party system, unlike the two-party system of the US Senate, the classification task at hand is more complex than the classification task preformed in previous political science applications of the SVM algorithm. For analytical purposes SVM applications divide the documents into a training set and a test set. In the training set the group membership is know. The task in the test set is to correctly predict the group membership of each document. In the email spam example, the training algorithm aims to correctly predict whether a new email (a new document) belongs to a spam or a non-spam category. The SVM model assumes that data points in each of the categories are separable by a hyperplane. SVM also assumes that there exists an "ideal" primary hyperplane lying at equal distances between two parallel secondary hyperplanes, each of which is determined by one or more data points in one of the assigned grouping categories. These data points on the parallel hyperplanes are called Support Vectors. The distance between the hyperplanes is called the margin. The task of the Support Vector Machine in the training phase is to find the separating hyperplanes in order to maximize the value of the margin. Our analysis is more complex because we are not simply classifying the documents into two categories (e.g. spam – not spam). Our aim is to classify political speeches by party group affiliation, which contains at least seven categories.

In the SVM text classification model, documents are represented in a vector space whose dimensions correspond to the features that are relevant for the classification task. In the following analysis, the relevant features are words spoken by legislators and the vector representing each document is determined by the number of occurrences of each of the words in that specific document. The documents consists of all the legislative speeches made by a specific legislator in a given European Parliament. We thus construct a $m \times n$ matrix where m the numbers of rows represent each individual MEPs, and n, the number of columns represent each individual words spoken in the legislature. The entry on the i^{th} row and the j^{th} column represents the number of times a specific word occurs in all of the speeches made by a particular legislator.

It is important to note that the proposed analysis will only make sense if we assume that speech can be constrained by ideology in the legislature. If we define ideology as a belief system which gives structure to an individual's view on various political issues, we can expect such belief systems to constrain individual positions on certain issues. In other words, "constraint may be taken to mean the success we would have in predicting, given an initial knowledge that an individual holds a special attitude, that he holds certain further ideas and attitudes" (Converse, 1964, 207). Thus, it is quite possible (though not guaranteed) that a randomly selected MEP legislator who opposes stricter environmental legislation also supports more flexible labor laws.

Not surprisingly, it is very difficult to measure ideological orientations and belief systems since it is not directly observable. In the U.S. Congress literature, the most widely used measure of ideology remains the vote-based score NOMINATE developed and refined by (Poole and Rosenthal, 1997). However, as was indicated earlier, we have chosen to use in this study a very basic measure of ideological belief system. We assume that party affiliation can serve as a proxy for ideological orientation in a, possible multidimensional, policy space. Carrubba et al. (2006) argue that selection bias in roll call votes may disguise some of the dimensions of conflict as some policy areas have relative very few roll call votes compared to the relative presence on the agenda in EP plenary sessions. Reliance on NOMINATE scores as a proxy of individual ideology may hence be problematic if the debates in the European Parliament is divided primarily along other lines that are not observed in roll call votes, as was shown to be the case in the 5^{th} EP (Proksch and Slapin, forthcoming).

This explains why we focus on party affiliation. If there is indeed a high level of vocabulary cohesion among party group members from different states (who speak a different language), we can assume like Diermeier et al. (2006) that speakers are talking about the same issues on the legislative floor. However, if we find that there is a very low classification success rate among members of the same party, we can reject the idea that legislators are making speeches about the same issues. Finding agreement evidence between different legislative speeches can be a powerful aid in an automated classification task (Thomas, Pang and Lee, 2006). We expect to find evidence of a high level of similarity between two speakers who share the same party group affiliation if both speakers share the same belief system (e.g. they will debate on the same issues, and utilize the same vocabulary).

DATA

We downloaded all the MEPs speeches of the 5^{th} and first half of the 6^{th} EP from the website `http://www.europarl.europa.eu`. We then converted the original html files to pure text by removing the html tags, headers, tables, lists, and non-acsii characters. We also segmented the speech files into individual legislator speeches. An individual file is all the words spoken by a particular MEP during the whole parliamentary term. We used party-group and country affiliation from the MEP biographic data as identifiers of party affiliation (Hoyland, Sircar and Hix, 2009). A training document is an MEPs complete set of speeches in the 5^{th} EP, and a test document is one's complete set of speeches in the first half of the 6^{th} EP. We exclude words that are used by less than 5% of the MEPs. The preparation of the data was done in R (R Development Core Team, 2008) with the *tm* package (Feinerer, Hornik and Meyer, 2008). We stemmed the terms, using the standard Porter stemmer.[1] We weight the stems by the inverse of the stems in the documents. We also excluded stop-words and words used by fewer than 5% of all MEPs. This gives us 4440 word-stems. We have training-sets for 702 MEPs from 5^{th} EP, and test documents for 752 MEPs in the first three years of the 6th EP. In the 6^{th} Parliament, there are test documents from 567 MEPs from the old member states and 185 test documents from the new member states. Of the members from the old member states, 300 of the MEPs are incumbents and 267 freshmen.

We use the *ksvm* function in the kernlab library (Karatzoglou et al., 2004) in R (R Development Core Team, 2008) for the classification tasks. As the classification task is

[1]Stemming is a process to remove the suffix of a word, for example "stemming" can be stemmed into "stem".

multi-category, we used the Weston and Watkins native multi-class classifier with a linear kernel function, as is common in text classification tasks where the information in the data-matrix is sparse. The *minpair* coupler was used to adjudicate between alternative groups (Wu, Lin and Weng, 2004).We set the cost parameter $C = .00005$. We do this in order not to over-fit the *Support Vector Machine* to the training-data. Adjustments in the cost parameter have small effects on the classification results. Such adjustments do not change the substantive results.

Members of the European Parliament represent national parties. These national parties form party groups with like-minded national parties from other member states. Most party groups have only one party from each country, but some exceptions exists (Corbett, Jacobs and Shackleton, 2000). The largest party group is the European Peoples Party and European Democrats (EPP-ED). National parties affiliated with this party tend to hold a Christian democratic or Conservatives ideology or belief system. The second largest party group is the Party of European Socialists, (PES). The national parties in this party group tend to share a social democratic ideology or belief system. The third party group is the liberal. During the 5^{th} Parliament it was named the European Liberal and Democratic Reformist party. Later it re-branded itself as the Alliance of Liberal and Democrats for Europe (ALDE) at the start of the 6^{th} Parliament. As the name suggest it consists of national parties that promote a liberal ideology or belief system. The Green / European Free Alliance (Green) party group consists of national parties whose key common denominator is a concern over the environment. The European United Left / Nordic Green Left (Left) consists of national parties with a socialist, new leftist, or communist belief system or ideology. Union of Europe of the Nations (UEN) caters for national parties that share a belief system or ideology that emphasizes the role of the national state. It is commonly placed to the right of EPP-ED on the political spectrum. There is also an independent democrats group and some independent members (ID). These are normally skeptical of the project of European integration, but members may hold political beliefs of both left and right standings. We use all the speeches deliver by members of these groups during the 5^{th} Parliament to train our classifier. The training error is .013. There are 652 relevant support vectors. We then use these to predict the party affiliation of the plenary speakers during the first three years of the 6^{th} Parliament. The results are reported in the next section.

RESULTS

This section evaluates the performance of the *Support Vector Machine* in correctly predicting the party group of plenary speakers in the 6^{th} European Parliament. Table 1 presents the overall results. Table 2 present the results for speakers from the old member states only. Table 3 shows the results for speakers from the new member states. Table 4 shows the predictions from freshman MEPs from the old member states. Finally, table 5 tabulates the predicted party group affiliation of incumbents against their actual party group affiliation.

Support Vector Machine Predictions of Party Group Affiliation all MEPS

	TRUE ALDE	Green	Left	ID	EPP-ED	PES	UEN	precision
ALDE	10	0	0	1	8	2	0	.48
Green	3	19	5	0	1	3	0	.61
Left	1	4	25	1	4	2	1	.66
ID	3	1	0	30	3	3	2	.71
EPP-ED	41	10	6	20	184	70	24	.52
PES	35	8	8	8	62	121	11	.48
UEN	1	0	0	1	7	2	1	.08
overall	94	42	44	61	269	203	39	752
recall	.11	.45	.57	.49	.68	.60	.03	
F-score	.17	.52	.61	.58	.59	.53	.04	
$\chi^2 \approx 38.893$	df= 1	$p \approx .00$						.52

Table 1: Cross-table of predicted versus true party group affiliations of plenary speakers as predicted by the Support Vector Machine. Training-data = European Parliament 1999 - 2004 (all MEPs). Test-data = European Parliament 2004 - 2007, (all MEPs).

Table 1 tabulates the predicted party group membership of the speakers against their actual membership. If the model fitted the data perfectly, all observations should be on the diagonal from upper left to lower right. Large deviations indicate poor fit. Overall, the *Support Vector Machine* correctly predicts the party group affiliation of .52 of all speakers of all speakers. Without any prior information except the proportion of the groups, the best we could do would be to guess that any randomly selected speaker belonged to the largest party group. This *majority rule* of always guessing the largest party group would be right $\frac{269}{752} \approx .36$. The χ^2 test rejects the hypothesis of equality in the proportion of correct predictions between the *Support Vector Machine* and the majority vote rule, $\chi^2 \approx 38.893$, $p \approx .00$. The performance of *Support Vector Machines* is often evaluated in terms of *precision* and *recall*. *Precision* is the percentage of proposed members of a group who is actually in that group. *Recall* is the percentage of actual members of the group correctly identified as members of that group. Ideally, we would like a high level of both prediction and recall. A common measure of this is the *F-score*, the harmonic mean of recall and precision. The harmonic mean give higher value to numbers that are close together than numbers that are far apart. Both *precision* and *recall* are reported in the table. The discussion in the text focus mainly on the *F-score*.

The *F-score* is above .50 for all party groups except the liberal ALDR and the rightist UEN. The *Support Vector Machine* makes many mistakes on recall of members of ALDR. The large majority of ALDR members are misplaced amongst the EPP-ED or as PES. Similar mistakes occur in terms of precision, the *Support Vector Machine* places several members of EPP-ED as members of ALDR. This is not surprising. The ALDR is ideologically located between these two party groups on the standard left-right dimension. The ALDR alternates between forming coalitions between the PES and EPP-ED in roll-call votes (Hix and Noury, 2006). The *Support Vector Machine* preforms best when it classifies MEPs the Left. The *F-score* is .61. It mis-classify several GUE/NGL members as Greens. Similarly, several Green MEPs are mis-classified as members of the Left. Again, this is not surprising. These two party groups have a common ideological stand on many issues. Their voting record is also very similar (Hix, Noury and Roland, 2007). The ability of the *Support Vector Machine* to recall members of the EPP-ED is .68. This is very good. However, the precision is not as impressive, only .52 of its members are actually classified in that group.

The *Support Vector Machine* places too many MEPs in the EPP-ED category. Nevertheless, the *F-score* is still .59. The algorithm is doing considerably better than simply guessing EPP-ED membership when we focus on all members of the European Parliament. Such a guess would of course reflect a perfect *recall*. But the precision would be below .39 mark. In this scenario, the *F-score* would be .56. Our analysis of *Support Vector Machine* is actually able to recall .60 of all MEPs from the PES. The precision is however only .48. Many PES members are classified as members of EPP-ED and vise-versa.

We have now demonstrated that the *Support Vector Machine* obtains acceptable classification results on the data from EP plenary debates. In the next subsection, we investigate whether there is any difference in the performance of the algorithm when we classify MEPs from old and new member states.

MEPs from old Member States

We begin by repeating the classification exercise for MEPs from the old member states only.

Support Vector Machine Predictions of Party Group Affiliation old Member States

	TRUE ALDE	Green	Left	ID	EPP-ED	PES	UEN	precision
ALDE	7	0	0	1	5	2	0	.47
Green	2	18	5	0	1	2	0	.64
Left	0	4	24	1	1	1	0	.77
ID	3	1	0	27	2	3	2	.71
EPP-ED	32	10	3	7	139	54	7	.55
PES	27	8	5	4	44	102	5	.52
UEN	1	0	0	1	5	0	1	.12
overall	72	41	37	41	197	164	15	567
recall	.10	.44	.65	.66	.71	.62	.07	
F-score	.16	.52	.71	.68	.62	.57	.09	
$\chi^2 \approx 51.225$	df= 1	$p \approx .00$						.56

Table 2: Cross-table of predicted versus true party group affiliations of plenary speakers as predicted by the Support Vector Machine. Training-data = European Parliament 1999 - 2004. Test-data = European Parliament 2004 - 2007, (only MEPs from old member states).

Table 2 tabulates the predicted party group membership of the speakers against their actual membership form MEPs of the old member states. The model predict the correct party group of .56 of all MEPs from the old member states. Naively guessing the largest party group for all observations would results in $\frac{197}{567} \approx .35$ correctly predicted. The χ^2 test rejects the hypothesis of equality in the proportion of correct predictions between the

Support Vector Machine and the majority vote rule,$\chi^2 \approx 51.225$, $p \approx .00$.

Again, we find that the *F-score* is above .5 for all party groups except the liberal ALDE and the rightist UEN. When compared to the previous classification task, the *F-scores* actually improve for all party groups.

Note that it reaches an impressive *F-score* of .71 for the left-wing GUE/NGL. Independents are also very well predicted. For this group, the *F-score* is .68. In addition, the *F-score* for the largest party group, EPP-ED is .62. The *recall* measure is also very good, .71. In other words, the *Support Vector Machine* is able to correctly classify more than 7 out of every 10 EPP-ED member. It does, however, slightly over-predict EPP-ED membership. Here, the *precision* is only .55. Most of the precision mistakes are made for members of the PES and ALDE. The *F-score* for the *Support Vector Machine* for the EPP-ED compares favorably to the strategy of always guessing EPP-ED, which would obtain an *F-score* of .52.

A similar picture emerges for PES. The *F-score* is .57., the *recall* measure is .62, and the *precision* measure is .52. In the next subsection, we report the classification results for MEPs from the new member states only.

MEPs from new Member States

Table 3 presents the results obtained when using the model to predict party group membership of MEPs from the new member states. As the table show, the fit is poor. Naively guessing the largest party group would result in a proportion correctly predicted of $\frac{72}{185} \approx .39$. The test statistic, $\chi^2 \approx 0$, $p \approx 1$, confirms that we cannot reject the hypothesis that the *Support Vector Machine* and majority vote are equally (ill-) suited for predicting party group membership amongst members from the new member states. The strategy of guessing that all MEPs from the new member states belong to the EPP-ED would result in an *F-score* of .56. This is better than the F-score of .51 obtained by the *Support Vector Machine*. This suggest that members from states who joined the EU in the 6^{th} Parliament do not share the same type of vocabulary as the one expressed by members of their party group in the 5^{th} Parliament. Based on this finding, it would appear that the choice of new MEPs to join party groups is not necessarily motivated by a similarity

Support Vector Machine Predictions of Party Group Affiliation new Member States

	TRUE ALDE	Green	Left	ID	EPP-ED	PES	UEN	precision
ALDE	3	0	0	0	3	0	0	.50
Green	1	1	0	0	0	1	0	.33
Left	1	0	1	0	3	1	1	.14
ID	0	0	0	3	1	0	0	.75
EPP-ED	9	0	3	13	45	16	17	.44
PES	8	0	3	4	18	19	6	.33
UEN	0	0	0	0	2	2	0	.00
overall	22	1	7	20	72	39	24	185
recall	.14	1	.14	.15	.62	.49	.00	
F-score	.21	.50	.14	.25	.51	.39	.00	
$\chi^2 \approx 0$	df= 1	p≈ 1						.39

Table 3: Cross-table of predicted versus true party group affiliations of plenary speakers as predicted by the Support Vector Machine. Training-data = European Parliament 1999 - 2004. Test-data = European Parliament 2004 - 2007, (only MEPs from new member states).

in their belief systems. If this where the case, we would find a greater complementarity in the content of speeches of both new and old members from the same party group. There may be several reasons which can explain this result. First, as we just indicated, it may be that MEPs from the new member states do not share the same belief system as the MEPs from the party group they choose to join. It may also be that the parties from the new member states choose to join a certain party on the basis of office spoils, rather than similarity in belief systems. It may be the case that although these MEPs share the same broad belief system held by the other MEPs in their party group, they may have not yet become familiar with the standard (Western European) ways of expressing these views (hence the poor classification). It may finally be the case that these similarities are lost in the process of translating the speech from their national language to English. However, Proksch and Slapin (forthcoming) obtain very similar results when running *Wordfish* on the German, French and English corpus of debates in the 5^{th} Parliament.

One should however bear in mind that it is quite possible that the model is better at predicting the group membership amongst MEPs from the old member states because some MEPs appear both in the training and the test data, as they were (speaking) members in both the 5^{th} and 6^{th} Parliaments. The increased prediction rate may thus be explained by the fact that the *Support Vector Machine* is recognizing individual members from both sets of data. In order to ensure that our results are not driven by this "incumbency effect", we also predicted the party group membership of freshmen originating from old member states only. By freshmen, we mean members who joined the EP after the 2004 elections. These results are presented in table 4. We also contrast theses results with the results obtained when the *Support Vector Machine* predicts party group affiliation in the 6^{th} Parliament amongst members appearing in both the training and the test data.

Support Vector Machine Predictions of Party Group Affiliation old Member States (freshmen)

	TRUE ALDE	Green	Left	ID	EPP-ED	PES	UEN	precision
ALDE	1	0	0	0	3	1	0	.20
Green	1	8	3	0	1	0	0	.62
Left	0	1	15	1	0	1	0	.83
ID	2	0	0	13	1	1	2	.68
EPP-ED	24	6	2	5	48	25	3	.42
PES	18	6	4	4	18	42	1	.45
UEN	1	0	0	1	3	0	1	.17
overall	47	21	24	24	74	70	7	267
recall	.02	.38	.62	.54	.65	.60	.14	
F-score	.04	.47	.71	.60	.51	.52	.15	
$\chi^2 \approx 22.367$	df= 1	$p \approx .00$						.48

Table 4: Cross-table of predicted versus true party group affiliations of plenary speakers as predicted by the Support Vector Machine. Training-data = European Parliament 2004 - 2007 (only freshmen from old member states). Test-data = European Parliament 1999 - 2004.

Table 4 tabulates the predicted party group membership of the speakers against their actual membership for freshmen originating from the old member states. Majority voting obtains a benchmark classification rate of $\frac{74}{267} \approx .28$. The support vector machine obtians a classification rate of .48. The improvement is significant, $\chi^2 \approx 22.367$, $p \approx .00$. The strategy of always guessing the biggest party group results in an *F-score* of .44. The *Support Vector Machine* obtains an *F-score* of .51 for the EPP-ED. This clearly shows that the difference in model performance on speeches delivered by MEPs from old versus new member states is not driven by incumbents appearing in both the training and the test data.

Support Vector Machine Predictions of Party Group Affiliation old Member States (incumbents)

	TRUE ALDE	Green	Left	ID	EPP-ED	PES	UEN	precision
ALDE	6	0	0	1	2	1	0	.60
Green	1	10	2	0	0	2	0	.67
Left	0	3	9	0	1	0	0	.69
ID	1	1	0	14	1	2	0	.74
EPP-ED	8	4	1	2	91	29	4	.65
PES	9	2	1	0	26	60	4	.59
UEN	0	0	0	0	2	0	0	.00
overall	25	20	13	17	123	94	8	300
recall	.24	.50	.69	.82	.74	.64	.00	
F-score	.34	.57	.69	.78	.69	.61	.00	
$\chi^2 \approx 29.095$	df= 1	$p \approx .00$						.63

Table 5: Cross-table of predicted versus true party group affiliations of plenary speakers as predicted by the Support Vector Machine. Training-data = European Parliament 2004 - 2007 (only incumbents from old member states). Test-data = European Parliament 1999 - 2004.

For completeness, table 5 presents the results for incumbent MEPs only. The correctly predicted party group membership is high, .63, the χ^2 test does allow us to reject the null hypothesis that majority vote perform equally well on this data, $\chi^2 \approx 29.095$ $p \approx .00$. Always predicting EPP-ED gives a classification rate of $\frac{123}{300} \approx .41$. The *F-score* is .58. This compares to .69. The *Support Vector Machine* in this context is clearly more efficient at classifying MEPs, as we would expect.

CONCLUSION

In this paper, we used Support Vector Machine, a common technique for semi-automatic classification of text in computational linguistics, to investigate whether it is possible to correctly predict party group affiliation of MEPs in the 6^{th} European Parliament by analyzing their speeches. This was done by training a SVM classifier on the legislative speeches made in the 5^{th} European Parliament to predict the party affiliation of elected representatives in the next parliament. The results show that while MEPs from the old member states express a belief system which is quite similar to that of their fellow party group members in the previous parliament, we found that MEPs from new member states displayed very little consistency in their legislative speech with former elected MEPs. This finding leads us to conclude that national party delegations from new member states joined the existing party groups for other reasons than simple shared ideological beliefs and goals.

REFERENCES

Baayen, R.H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.* Cambridge, Cambridge University Press.

Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara and Eric Tanenbaum. 2001. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-1998.* Oxford: Oxford University Press.

Carrubba, Clifford J., Matthew Gabel, Lacey Murrah, Ryan Clough, Elizabeth Montgomery and Rebecca Schambach. 2006. "Off the Record: Unrecorded Legislative Votes, Selection Bias and Roll-Call Analysis." *British Journal of Political Science* 36(11):691–704.

Clegg, Nick and Michiel van Hulten. 2003. Reforming the European Parliament. Technical report London: .

Converse, P. E. 1964. The Nature of Belief Systems in Mass Publics. In *Ideology and Discontent.* New York: Free Press.

Corbett, Richard, Francis Jacobs and Michael Shackleton. 2000. *The European Parliament.* 4th ed. London: John Harper.

Cousins, Ken and Wayne McIntosh. 2005. "than Typewriters, More than Adding Machines: Integrating Information Technology into Political Research." *Quality and Quantity* 39:591 – 614.

Diermeier, Daniel, Jean-François Godbout, Bei Yu and Stefan Kaufmann. 2006. "Language and Ideology in Congress.".

Eriksen, Erik Oddvar and John Erik Fossum. 2000. *Democracy in the European Union - Integration through deliberation?* London: Routledge.

Eriksen, Erik Oddvar and John Erik Fossum. 2002. "Democracy through Strong Publics in the European Union." *Journal of Common Market Studies* 40(3):401–24.

Farrell, David, Simon Hix, Mark Johnson and Roger Scully. 2006. "EPRG 2000 and 2006 MEP Surveys Dataset.".
URL: *http://www.lse.ac.uk/collections/EPRG/survey.htm*

Feinerer, Ingo, Kurt Hornik and David Meyer. 2008. "Text Mining Infrastructure in R." *Journal of Statistical Software* 25(5):1–54.

Gabel, Matthew and Simon Hix. 2007. "From Preferences to Behaviour: Comparing MEPs' Survey Response and Roll-Call Voting Behaviour." Paper prepared for presentation at the Tenth Biennial Conference of the European Union Studies Association, Montreal, Canada, 17-19 May 2007.

Hix, Simon. 2002. "Parliamentary Behavior with Two Principals: Preferences, Parties, and Voting in the European Parliament." *American Journal of Political Science* 46(3):688–689.

Hix, Simon. 2004. "Electoral Institutions and Legislative Behavior: Explaining Voting-Defection in the European Parliament." *World Politics* 56(1):194–223.

Hix, Simon and Abdul Noury. 2006. "After enlargement: Voting Patterns in the Sixth European Parliament." Memio, London School of Economics.

Hix, Simon, Abdul Noury and Gerard Roland. 2005. "Power to the Parties: Cohesion and Competition in the European Parliament, 1979-2001." *British Journal of Political Science* 35(2):209–234.

Hix, Simon, Abdul Noury and Gerard Roland. 2007. *Democracy in the European Parliament*. Cambridge University Press.

Hopkins, Daniel and Gary King. 2007. "Extracting Systematic Social Science Meaning from Text." Unpublished Manuscript, Harvard University.

Hoyland, Bjorn, Indraneel Sircar and Simon Hix. 2009. "An Automated Database of the Europea Parliament." *European Union Politics* 10(1).

Karatzoglou, Alexandros, Alex Smola, Kurt Hornik and Achim Zeileis. 2004. "kernlab – An S4 Package for Kernel Methods in R." *Journal of Statistical Software* 11(9):1–20. **URL:** *http://www.jstatsoft.org/v11/i09/*

Klingemann, Hans Dieter, Andrea Volkens, Judith Bara and Michael Budge, Ian andM-cDonald. 2006. *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990 - 2003.* Oxford: Oxford University Press.

Laver, Michael and John Garry. 2000. "Estimating Policy Positions from Political Texts." *American Journal of Political Science* 44(3):619–634.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97(2):311–331.

Mamadouh, Virginie and Tapio Raunio. 2003. "The Committee System: Powers, Appointments and Report Allocation." *Journal of Common Market Studies* 41(2):333–351.

Monroe, Burt L. and Ko Maeda. 2004. "Rhetorical Ideal Point Estimation: Mapping Legislative Speech." Society for Political Methodology, Standfor University, Palo Alto.

Poole, Keith T. and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll-Call Voting.* Oxford: Oxford University Press.

Proksch, Sven O. and Jonathan B. Slapin. forthcoming. "Position Taking in European Parliament Speeches." *British Journal of Political Science* .

Purpura, S and D. Hillard. 2006. Automated Classification of Congressional Legislation. In *Proceedings of Digital Government Research.*

Quinn, Keven M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin and Dragomir R. Radev. 2006. "An Automated Method of Topic-Coding Legislative Speech Over Time with Appliucation to the 105^{th} - 108^{th} U.S. Senate.".

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-9000051-07-0. **URL:** *http://www.r-project.org*

Schonhardt-Bailey, Cheryl. 2006. *From the Corn Laws to Free Trade: Interests, Ideas, and Institutions in Historical Perspective*. Cambridge: MIT Press.

Slapin, Jonathan B. and Sven O. Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705 – 22.

Thomas, Matt, Bo B Pang and Lillian L. Lee. 2006. Get out the vote: DEterminging support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*. EMNLP pp. 327 – 335.

Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. New York: Wiley.

Vapnik, Vladimir N. 1999. *The Nature of Statistical Learning Theory*. 2 ed. New York: Springer.

Wu, Ting-Fan, Chih-Jen Lin and Ruby C. Weng. 2004. "Probability Estimates for Mulit-class Classification by Pairwise Coupling." *Journal of Machine Learning Research* 5:975 – 1005.