Applying Machine Learning to Political Discourse (draft)

Analyzing Parliamentary Discourse on Educational Reform (Belgium, 1879, France 1880)

Kaspar Beelen (University of Antwerp), **Marnix Beyen** (University of Antwerp)

Introduction

Machine Learning

Recently the digitization of parliamentary documents increased enormously. Not only contemporary proceedings were published online, historical debates too are gradually becoming accessible through the internet. The Belgian parliament is trying not to lag behind in this respect. For several years photographed copies of the proceedings could be found on their website The project '*plenum.be*', very recently initiated further digitization of these documents and aims to disclose all debates as searchable text this year.

Digitization could serve as a new stimulus for the study of parliamentary history. The proceedings are now easily available for everyone and in some tenths of a second the historian can trace his or her particular fetish. Thank you Google. But how to cope with these huge libraries? More specifically: how can we use these databases for *historical* research? How to study parliamentary rhetoric in all its immensity and detect structural patterns in this unwieldy discursive mass, patterns that are both statistically and historically relevant? In this paper I'll gradually develop a method that will satisfy both conditions, I hope. To accomplish this goal I will apply data mining techniques and machine learning to parliamentary discourse.[1]

I won't spend too much time on explaining how machine learning exactly works. By giving a straightforward example I will merely define its basic characteristics and possibilities. I'll concentrate on a specific kind of learning (supervised classification), and a specific application (spam filtering). Classification amounts to ascribing labels to documents, labels like 'spam' and 'non-spam'. Because all the documents are labeled manually according to their '*class*' beforehand, the learning is 'supervised'.

First we have to define the linguistic characteristics or '*features*' in which we are interested, and extract them from the documents. An example of features are words and their frequencies. In more formal terms, we assign every document or '*instance*'(w^j)to a class and reduce it to a vector with x_i de features (words) and a_i its values (frequencies).

Afterwards de database is split into a train set and a test set. By training, the learner is able to scrutinize relations between features and classes. It will probably find out that the word 'Casino' or '100.000.000$' are more likely to appear in documents categorized as spam. The important features for distinguishing texts are called here the 'most informative features (MIF)'. The model produced by training will be applied to a test set. For each instance in the test set a label is calculated and then compared to the original class. This allows for quantifying an accuracy. Accuracy is the ratio between the correctly classified documents in the test set to all the documents in the test set. It indicates how well the classifier is in predicting the label of a text. The higher the accuracy the more the classifier is able to correctly identify 'spam' and 'non spam' mails by scrutinizing the linguistics characteristics of a mail (in this examples word frequencies). In this lecture I will apply the same principles to political rhetoric and search for discursive differences between groups in parliament.[2]

---

[1] <http://www.plenum.be> and <http://hansard.millbanksystems.com>, <http://www.statengeneraaldigitaal.nl>

[2] For Classification I used: 'Orange' <http://www.ailab.si/orange/>, 'Weka3.6': <http://www.cs.waikato.ac.nz/ml/weka/> or Classification module in the NLTK toolkit(NLTK handbook, chapter 6: <http://nltk.googlecode.com/svn/trunk/doc/book/ch06.html>)<http://nltk.googlecode.com/svn/trunk/doc/api/nltk.classify-module.html>. Witten, J & Frank, E., '*Data Mining : Practical Machine Learning Tools and Techniques*', Morgan Kauffman Publishers, 2005, p. 97 – 105. For pre-processing Python was used: Lutz, M., '*Learning Python*', O'Reilly, 2008; Hetland, M., Segaran, T, '*Programming Collective Intelligence*', O'Reilly, 2007; Hetland, M., '*Beginning Python from novice to professional*', Apress, 2008. Jurafsky, D. & Martin, J., '*Speech and Language Processing: An Introduction to*

The models I'll apply below, fit within the framework of 'Natural Language Processing'[3], i.e. a computational approach to human or 'natural' language. Trained as a historian, I was totally ignorant of computational linguistics. Throughout my academic career. I acquired no specific computer skills. With this paper I hope to prove that, being far from a specialist, applying computational techniques to political discourse isn't that hard in the end, even for a layman. I surely don't possess the qualities of a computer scientist, only those of a hobbyist who is able to tinker at lines of code.[4] Since I don't consider myself a specialist, this paper is also written for an audience without any foreknowledge of Machine Learning or Artificial Intelligence. All technical details and confusingly detailed tables are left out.

The programming Historian

My aim here is to differentiate between collectivities in parliament on the basis of their speeches. Would a machine be able to recognize the party political ideology of a speaker automatically? Hereby assuming that ideology is contained, or at least reflected, in language use. Do rightwing MPs speak the same language as those representatives occupying the benches on the left? If true, where to place the linguistic divide? How wide is the gap?

Machine learning has many applications in authorship attribution, like distinguishing Shakespeare's plays from Marlowe's. Beside comparing individuals, some studies investigated groups, and differentiated between male or female authors, age groups or ethnicities.[5] I opted for analyzing dissimilarities between political parties. First and foremost I was inspired by Prost's '*Vocabulaires des proclamations électorales*'. In this work he studied word frequencies in French '*professions de foi électorales*' at the end of the nineteenth century.[6] By analyzing vocabulary he tried to expose differences between left- and rightwing discourse.

Natural Language Processing, Computational Linguistic, and Speech Recognition', Prentice Hall, 2000. Bird, S. Klein, E., & Loper, E., '*Natural Language Processing With Python: Analyzing Text With Natural Language Toolkit*', O'Reilly, 2009.

[3] Jurafsky, D. & Martin, J., '*Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistic, and Speech Recognition*', Prentice Hall, 2000. Bird, S. Klein, E., & Loper, E., '*Natural Language Processing With Python: Analyzing Text With Natural Language Toolkit*', O'Reilly, 2009.

[4] I concentrated on using Python : Lutz, M., '*Learning Python*', O'Reilly, 2008; Hetland, M., Segaran, T, '*Programming Collective Intelligence*', O'Reilly, 2007; Hetland, M., '*Beginning Python from novice to professional*', Apress, 2008.

[5] Foster, D. 'The Claremont Shakespeare Authorship Clinic: How Severe Are the Problems?', *Computers and the Humanities*, 32:6, 1999. And Hirst G. & Feiguina, O., 'Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts', Literary and Linguistic Computing, 22:4, 2007, pp. 405 - 417.Argomon, S. & Olsen M., 'Words, Patterns and Documents: Experiments in Machine Learning and Text Analysis', *Digital Humanities Quarterly*, 3:2, 2009,. Goulain, J-B, Argomon, S. et.al., 'Vive la Différence! Text Mining Gender Difference in French Literature '*Digital Humanities Quarterly*, 3:2, 2009. Meehan, S., 'Text Minding: A Response to Gender, Race, and Nationality in Black Drama, 1850-2000: Mining Differences in Language Use in Authors and their Characters', *Digital Humanities Quarterly*, 3:2, 2009

[6] Prost, A. '*Vocabulaire des Proclamations Électorales de 1881, 1885 et 1889*' in : Travaux de centre de recherches sur l'histoire du XIXe siècle, 2, Presses Universitaires de France, 1974. Also see : Maldidier, D. & Robin D., 'Polémique idéologique et affrontement discursif en 1776: Les grands édits de Turgot et les remontrances du parlement de Paris' in : Robin, R., Guilhaumou, J., Maldidier, D. & Post, A. '*Langage et Idélogies : le discours comme objet de l'Histoire*', Les Éditions Ouvrières, 1974, pp. 13 80, Guilhaumou, J., 'L'Idéologie du Père Duchesne : les forces adjuvantes', in : Robin, R., Guilhaumou, J., Maldidier, D. & Post, A. '*Langage et Idélogies : le discours comme objet de l'Histoire*', Les Éditions Ouvrières, 1974, pp. 81 – 116. Prost, A. 'Combattants et Politiciens. Le discours mythologique sur la politique entre les deux guerres', in : Robin, R., Guilhaumou, J., Maldidier, D. & Post, A. '*Langage et Idélogies : le discours comme objet de l'Histoire*', Les Éditions Ouvrières, 1974, pp. 117 – 150.

His exclusive focus on singular words, didn't seem to serve my purposes. I will analyze the relationship between collocational restraints and political ideology.[7] Instead of scrutinizing vocabulary, I'm more interested in the (re)occurrence of textual patterns. The simplest examples of such patterns are '*n*-grams' or sequences of *n* (characters or) words. Nevertheless, only scrutinizing *n*-grams could be criticized as superficial. To further penetrate the structure of parliamentary language, the analysis is placed on a more syntactic level, by searching for regularities in the grammatical structure. The words itself and the syntactic links between word classes become paramount. Features are 'grammars' or clusters of word classes, like couples of adjectives and nouns. Special attention went to the syntactic use of person deixis, combinations with personal and possessive pronouns.[8]
This methodology will be tested on a relatively small corpus of around half a million words.
The text material is primarily drawn from debates on the reform of primary education (1879). The principal goal of the liberal government, headed by Frère-Orban and Van Humbeeck, was to loosen the clerical grip on primary education. During this period, a predominantly philosophical divide between left-wing Liberals and right-wing Catholics cuts through the Belgian parliament. Most of the calculations below are based on these debates.

Results
What are the results? Let's first examine the *n*-grams[9], i.e. sequences of *n* words. Trigrams, sequences of three words, seem to deliver the best results with an average accuracy of 80%. If we take a closer look at the ten most informative trigrams we get the following results (ranked in descending order of importance per class):

Trigrams (selection 1879)
accuracy[10] = 80%
MIF[11] **Catholic** : *le honorable rapporteur, au nom de, monsieur le ministre, de la gauche, de la majorité, de la chambre, de la liberté*
MIF Liberal : *de la droite, de nos adversaires*

Constructions consisting of '*de la*/*le*' combined with a noun are very effective in categorizing the data.[12] The debates of 1879 seem to be permeated with an ever present conflict. The 'adversarial'

---

[7] Bayley, P. 'Lexis in British Parliamentary Debate: Collocation Patterns', in: *Language and Ideology. Selected Papers from the 6th International Pragmatics Conference*, International Pragmatics Association, 1999, pp. 43 – 55.
[8] Maitland K. & Wilson, J., 'Pronominal Selection and Ideological Conflict', Journal of Pragmatics, 11:4, 1987, pp. 495-512, Wilson, J., *Politically Speaking, The Pragmatic Analysis of Political Language*, Blackwell, Oxford, 1990. Verschueren, J. '*Understanding Pragmatics*', Oxford University Press,, 1999, p. 20; Chilton, P. 'Analyzing Political Discourse: Theory and Practice', Routledge, 2003, p. (?).Maitland K. & Wilson, J., 'Pronominal Selection and Ideological Conflict', Journal of Pragmatics, 11:4, 1987, pp. 495-512, Wilson, J., *Politically Speaking, The Pragmatic Analysis of Political Language*, Blackwell, Oxford, 1990. Benoît, R., 'Les figures du parti : formation et définition du groupe (1932 – 1946)', *Mots*, 10, 1985, pp. 109 – 132.
[9] *n*-Gram-based classification is often for authorship attribution: Soboroff, I., Nicholas, C., Kukla, J. & Ebert, D., 'Visualizing Document Authorship Using N-grams and Latent Semantic Indexing', in: *Proceedings of the Workshop on New Paradigms in Information Visualization and Manipulation*, 1998 pp. 43-48; Luyckx, K, '*Syntax-Based Features and Machine Learning techniques for Authorship Attribution*', Unpublished Master Thesis, 2004, p. 10. < http://www.clips.ua.ac.be/~kim/Papers/MAThesis_KimLuyckx.pdf> It even has biomedical applications: . Tomović, A, Janičić, P. & Kešelj, V., 'n-Gram-based classification and unsupervised hierarchical clustering of genome sequences', *Computer Methods and Programs in Biomedicine*, 81:2, 2006, pp. 137 – 153; And Hirst G. & Feiguina, O., '*Bigrams of Syntactic Labels*'.
[10] Classifier: Naïve Bayes; validation: 10 times repeated randomization of feature sets: train set, test set = 50%, lemmatization=no
[11]

nature is reflected in constantly talking about the Other. '*Monsieur le ministre*' '*de la gauche* 'and '*de la majorité*' typify the discours of the right-wing opposition, while Liberals principally refer to '*de nos adversaires*' and '*de la droite*'. This might seem a rather trivial result , but is this true? Also in other cases, the classifier consistently points to the importance of adversarial language. Speeches are best differentiated by more formal aspects of parliamentary rhetoric, such as official titles. Catholics focus on the liberal Cabinet ('*honorable ministre*') while their political opponents indirectly address their opponents ('*honorable monsieur*' or '*honorable membre'* (table 4 below)). The assertion concerning adversarial discourse could find further support when looking at nouns or names[13] and combinations of nouns or names with at least one adjective[14]. A selection from the thirty most important features leads to following results:

Grammar = {<NOM|NAM>}(selection) (1879)
Accuracy[15] = 76%
MIF Catholic : é*cole, moralité, liberté, commune, état, enfant, olin, éducation, loge, rapporteur, constitution*
MIF Liberal :clergé, église, *autorité, malou, prêtre, évêque, opposition, loi,woeste*

The names 'Malou', 'Woeste' and 'Beernaert' all rightwing MPs, are more likely to appear in liberal speeches, while representatives on the right regularly drop the names of their left wing adversaries 'Van Humbeeck' and 'Olin', or they refer to '*ministre*' or '*rapporteur*', functions that at the time were met by liberals. The same is true for other nouns. The left focuses on subversive activities of the clergy ('*clergé*') supported by the Catholic parliamentarian opposition ('*opposition*'). '*Clergé*' occupies the first place among the liberal distinctive features, followed by '*église*', '*prêtre*' and '*évêque*'. On the other half of the parliamentarian hemisphere, the State ('*État*'), the free masons ('*loge*') and the liberal majority ('*majorité*') have to suffer discursive attacks. The discourses on both sides principally differ in the characterization of the Other. *Hétérodésignation* rather then *autodésignation* defines this kind parliamentary discourse**.**
Of course MPs don't only talk about the Other. '*Liberté*', '*commune*' and '*dieu*' are the right's holy trinity. Catholics present themselves as passionate advocates of decentralization and guardians of the municipal powers. They defend municipalities against a State who's growing power threatens to monopolize society ('*monopole*'). Furthermore, there are some small lexical differences, such as the Liberal preference for '*foi*' instead of '*réligion*', or more symptomatic in this context, their choice for '*enseignement*' were Catholics opt for terms like '*école*', '*education*', or '*enfant*'. Maybe the Catholic support for the delegation of power to lower institutional levels, is connected with the a preference for concrete elements.

grammar ={(<ADJ>+<NOM|NAM><ADJ>*
{(<ADJ>*<NOM|NAM><ADJ>+)} (selection) (1879
accuracy[16] = 74%
MIF Catholic :*honorable ministre*, *instruction publique*, *section central*, *comité scolaire*, *école libre*, *morale chrétien*, *morale universel*, *défense national*, *état libéral*, *système anglais*, *morale moyen*
MIF Liberal: *honorable monsieur*, *honorable membre*, *autorité civil*, *parti libéral*, *opinion libéral, pouvoir public, croyance religieux*, *parti catholique*, *clergé catholique*

---

[12] Prost also notes the importance of partitive constructions : Prost, A., 'Combattants et Politicians. Le Discourse Mythologique sur la Politique entre Les Deux Guerres' in : Robin, R.(ed.), '*Language et Idéologies : le Discours comme objet de l'Histoire*', Les Éditions Ouvrières, 1974, pp. 117 – 150.
[13] grammar = {<NOM|NAM>}
[14]grammar ={(<ADJ>+<NOM|NAM><ADJ>*
{(<ADJ>*<NOM|NAM><ADJ>+)}
[15] Classifier = Multinominal Naïve Bayes, validation= 10-fold cross-validation
[16] Classifier = Multinominal Naïve Bayes, validation= 10-fold cross-validation, lemmatization=yes

Both '*parti catholique*' and '*parti libéral*' belong to a left-wing discourse. Liberals are less reluctant to affirm the existence of party formations, unlike Catholics they're not averse towards identifying themselves with their own party or movement ('*opinion libéral*'). '*Parti catholique*' is principally a liberal concept which their right-wing opponents seldom use, they sparsely drop the term '*parti conservateur*'. The governmental powers, according to liberals, belong to '[le] *autorité civil*' and' '[le] *pouvoir publique*', elements that contrast with the religious usurpers in the guise of '*croyance religieux*' or '*clergé catholique*'.

Note that Catholics don't emphasize religion but morality. They are the proponents of a '*morale chrétienne*' or '*morale moyenne*' and reject the possibility of a '*morale universelle*' a morality independent from any positive religion. The abstract ethical system which the left promotes, tends to harm the youth's conscience. This moral teaches things contrary to what they traditionally learn at home, thus creating a kind of moral dissonance.

'*État liberal*' has a negative connotation, because it underlines the State's partisan nature. The government headed by Frère-Orban transfigures the State into a party instrument. The monopoly is in the hands of only a small group of liberals, who will impose their will on a majority of Catholics. In other words, the new policy heralds the end of freedom of conscience or freedom *tout court*. The coupling of '*Etat*' with a loss of freedom is also reflected in the contrast between State Schools ('*école de l'État*') and '*école libre*', which coincides with '*école catholique*'.

Table 1 : Confusion Matrix,grammar = {<NOM|NAM>} (1879)[17]

Liberal	**Catholic**	**← *classified as***
159	32	**Liberal**
30	161	**Catholic**

Table 2: Confusion Matrix, grammar = {<ADJ>} (1879)

Liberal	**Catholic**	**← *classified as***
125	66	**Liberal**
39	152	**Catholic**

Table 3 : Confusion Matrix, grammar = {(<ADJ>+<NOM|NAM><ADJ>*)}
{(<ADJ>*<NOM|NAM><ADJ>+)} (1879)

Liberal	**Catholic**	**← *classified as***
135	56	**Liberal**
34	157	**Catholic**

One last important comment. These and other confusion matrices[18] show how the classifier consistently has more difficulties with recognizing Liberal than Catholic documents. It classifies more Liberal speeches as Catholic than the other way around. The Catholic discourse seems to be more coherent. This could be a historically interesting result. Keeping in mind the later collapse of the Liberal party, these matrices maybe point to an earlier discursive disintegration. More extensive research is needed here though.

Exploring Person Deixis

The adversarial discourse that underlies the debates, could be reflected in the use of person deixis. To point the differences in identification and alterization I investigated the use of the possessive pronouns '*notre*' and '*votre*'. Accuracy scored consistently low. But still some interesting divergences appear.

[17] Classifier = Multinominal Naïve Bayes, validation= 10-fold cross-validation, lemmatization=yes

[18] Results are confirmed by a classification using each word class separately.

grammar = {<NOTRE|NOS><NOM>} (selection) (1879)
accuracy[19] = 52%
MIF Catholic : *nos enfants, nos populations, notre pays, nos croyances, nos ancêtres*
MIF Liberal : *nos adversaires, nos institutions, nos écoles, nos principes, nos lois*

What is 'ours'? Liberals focus on institutions and principles, while Catholics concentrate more on specific social entities. For example the fact that one speaks of '*nos écoles*' where the other prefers '*nos enfants*' emphasizes the different orientations. By comparing other expressions, the same discrepancy comes to the surface. '*Nos principes*', '*nos institutions*', '*nos adversaires*' all frame within the left's political-institutional discourse. Catholics place other elements to the fore like '*nos ancêtres*', '*nos croyances*', '*nos populations*' and '*notre pays*' Not the institutions are paramount, but the more concrete communities.

grammar = {<VOTRE|VOS><NOM>} (selection) (1879)
accuracy[20] = 57%
MIF Catholic : *votre loi, vos instituteurs, vos rangs*
MIF Liberal : *votre système, votre parti, vos amis, vos mains, vos enfants, votre droit*

Combinations with '*vos*' and '*votre*' achieve a higher average accuracy of about five percent. In general this result is determined by the virulent Catholic dislike of '*votre loi*'. They no longer wish to participate in the legislative work of the Chamber, or so it seems. The bill is a liberal creation, where Catholics distances themselves from as far as possible. Additionally '*vos*' points to a collectivity sitting on the opposite benches. Opponents are mainly addressed as a group with phrases like '*votre parti*' or '*vos amis*' in the liberal discourse and '*vos rangs*' among Catholics.
Beside knowing what belongs to 'us' or 'you', it is important to find out what 'we' and 'you' are doing, i.e. the actions of the in- and outgroup. By analyzing the first en second person, it is possible to map this these actions. All verbs are shown in their lemmatized form, therefore abstraction is made of tense. Crucial here is whether verbs are used negatively or positively. Between liking and not liking is a huge gap I suppose, so certain adverbs like'pas' or 'jamais' etc..are included in the analysis.

grammar = {<NOUS><VER.+><ADV>} (selection) (1879)
accuracy[21] = 55%
MIF Catholic: *nous devoir, nous croire, nous demander, nous vouloir que, nous croire que*
MIF Liberal: *nous vouloir pas, nous proposer, nous défendre, nous pouvoir*

Among the verbs prominent in Catholic discourse, '*nous devons*', is followed by '*nous croyons*' and '*nous demandons*'. A possible interpretation could be the tension between opposition and majority. Rightwing MPs rather 'ask' than 'propose'('*nous proposons*'). These demands are related to an obligation ('*nous devons*') and a 'belief' ('*nous croyons (que)*'). The liberal discourse revolves around '*nous pouvons*' and '*nous voulons*'. Their task is to protect ('*nous défendons*') the Belgian institutions against the subversive activities of the clergy and their parliamentary fifth column.

grammar = {<VOUS><VER.+>}
{<VOUS><VER.+><ADV>*}

[19] Classifier: Naïve Bayes; validation: 10 times repeated randomization of feature sets: train set, test set = 50%, lemmatization=no
[20] Classifier: Naïve Bayes; validation: 10 times repeated randomization of feature sets: train set, test set = 50%, lemmatization=no
[21] Classifier: Naïve Bayes; validation: 10 times repeated randomization of feature sets: train set, test set = 50%, lemmatization=yes

accuracy[22] = 55%
MIF Catholic: *vous dire, vous mettre, vous tenir, vous prendre, vous savoir que*
MIF Liberal: *vous entendre, vous prétendre, vous admettre, vous invoquer, vous parler, vous espérer*

The second person reflects perhaps the same divergences in verb use between opposition and majority. Catholics should especially listen and understand ('*vous entendez*') while liberals mostly speak ('*vous dites*').Note that the left opts for '*vous parlez*' instead of '*vous dites*', a semantic nuance for which I currently have no explanation. Furthermore, liberals believe that their opponents don't act sincerely ('*vous pretendez*') and force them to make concessions ('*vous admettez*'). '*Vous tenez*' and '*vous prenez*' – characterizing a rightwing discourse – possibly refers to the supposedly liberal tendency to continually appropriate and keep for themselves. Politics is a game of taking and retaining.
Until now the ideological contrast between left and right prevailed, a somehow arbitrary choice of course. Perhaps the biggest discursive differences exist between various parliamentary generations and not between parties? To scrutinize this claim, for every speech the date of birth of the MP was added to the corpus. To obtain a uniform distribution the MPs were divided into two classes, one containing representatives born before 1827, the other consisting of all those born later. A quick test with different syntax-based patterns gave a negative result.[23] The classifier scored consistently lower when compared to earlier results. At present it is difficult to know whether this depends on the small size of the database, or that indeed party trumps age. Still, what remains an interesting question is if generations differ in their use of the pronominal system. Maybe older MPs, prefer to represent themselves more as independent individuals, speaking in the first person singular, while younger generations speak on behalf of others, using 'we' instead.
Although a rather low accuracy, doesn't allow us to be really confident of our conclusions, the use of personal pronouns among older MPs showed a tendency towards the first person singular as opposed to younger representatives, who more frequently spoke in the first person plural. Replacing date of birth by year of entry in Parliament, gave to a similar result. More research over a longer period and on different kinds of debates is needed however.
Machine learning is not limited to scrutinizing discursive divergences between groups, it also allows to uncover structural changes over time. Here I compare the debates on educational reform (1879) with parliamentary discussion on universal suffrage (1893) more than ten years later. Classifying noun phrases containing '*notre*' results in following output:

grammar = {<NOTRE><ADJ>*<NOM>+<ADJ>*} (1879 -1893)(selection)
accuracy[24] = 65 %
MIF 1879 : *notre adversaire, notre école, notre liberté, notre enfant, notre/NOTRE institution national, notre population, notre constitution, notre nationalité*
MIF 1893 : *notre collègue, notre honorable collègue, notre pays, notre programme, notre principe*

The shift from 'our opponents' ('*notre adversaire*') to 'our colleagues' ('*notre* (*honorable*) *collègue*'). is the most striking trend. In 1893 the emphasis is less on conflict than on group coherence. Perhaps because in this period the old opposition between left and right is no longer as obvious as during the 'School War'. The fact that '*notre école*' and '*notre enfant*' typify the discourse of 1879 is not surprising. To achieve an extension of voting rights a constitutional revision was needed. Although this wasn't the case in 1879, the phrase '*notre constitution*' characterizes the earlier debates. The constitution was for both parties of key importance to guarantee the survival of 'our' freedoms ('*notre liberté*'), 'our' institutions ('*notre institution national*') and 'our' national identity ('*notre*

[22] Classifier: Naïve Bayes; validation: 10 times repeated randomization of feature sets: train set, test set = 50%, lemmatization=yes
[23] Classifier = Multinominal Naïve Bayes, validation= 10-fold cross-validation, lemmatization=yes
[24] Classifier = Multinominal Naïve Bayes, validation= 10-fold cross-validation, lemmatization=yes

nationalité'). The identification with these elements was less dominant during the later discussions, where mainly the expression '*notre pays*' was used.

grammar = {<VOTRE><ADJ>*<NOM><ADJ>*}
Accuracy[25] = 54%
MIF 1879 : *votre loi, votre projet, votre droit, votre propre parti, votre ami*
MIF 1893 : *votre système, votre discours, votre part, votre principe*

What about changes in Othering? It's hard to draw any conclusion from this list. The accuracy scores poorly maybe because the classifier has less material at its disposal. Striking is that the syntactic combinations with the lemma '*votre*' decrease with approximately 40%, while expressions containing '*nos*' or '*notre*' increase with 25%, as if the parliamentary discourse shifts from the ingroup to the outgroup, a trend which I already noted before.

4.3 Transnational Classification
Because of difficulties with digitizing French parliamentary debates, the experiments transnational classification remain rather limited and very provisional. Here I compared some French debates on compulsory primary education (1880 – '81) with those on educational reform in Belgium (1879).

grammar = {(<ADJ>+<NOM|NAM><ADJ>*)}
{(<ADJ>*<NOM|NAM><ADJ>+)}[26]
accuracy[27] = 87%
MIF France: *instruction primaire, instruction religieux, contrainte légal, enseignement religieux, école publique, ancien régime, révolution français*
MIF Belgium: *conseil communal, section central, école communal, école normal, comité scolaire, défense national, parti libéral, administration communal*

Notwithstanding the small database, it's possible to distinguish topics in quite similar debates. Looking to the table one could infer different oppositions. In France there is an opposition between regimes, between '*ancien régime*' and the period after '[la] *révolution française*', while in Belgium the contrast between the State and the municipalities seems to predominate the discussion on primary education ('*comité scolaire*' opposed to '*conseil communal*', '*administration communale*'). Classifying phrases comprising possessive pronouns didn't give any satisfying results. Only the expressions '*notre caractère*' and '*votre église*' mildly typify the French parliamentary discourse. '*Notre liberté*' and '*notre institution*' tend to characterize the speeches of Belgian MPs.

Conclusion
To conclude I propose two reasons for applying machine learning to parliamentary discourse. One is the availability of immense digital libraries and the ability of machines to find structural patterns in these huge datasets. A second reason is a learner's capability of dealing with text on a more abstract level, like studying the correlation between textual and non textual elements. The provisional results drawn from a rather limited corpus, mainly point to the adversarial character of parliamentary rhetoric, although some indicators marked a shift towards promoting ingroup coherence in the later period.

[25] Classifier = Multinominal Naïve Bayes, validation= 10-fold cross-validation, lemmatization=yes
[26] Evaluation method = 10-fold cross-validation; Classifier = MultiNominal Naïve Bayes
[27] Classifier = Multinominal Naïve Bayes, validation= 10-fold cross-validation, lemmatization=yes

But everything written here is merely a beginning. This methodology asks for further improvement. Especially a lot of work has to be done on information extraction, on relating entities in sentences. Syntax-based classification is a good starting point, but it still I'm still far removed from the endpoint, if there is one. Furthermore, I want to experiment with more advanced classification algorithms like Support Vector Machines, although there exists a unofficial rule in data mining that more complex algorithms don't necessarily deliver better results. Still, this asks for further research.