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

Applying Machine Learning to Political Discourse: Theory and Practice (draft)

Analyzing Parliamentary Discourse on Educational Reform (Belgium, 1879)

1. Introduction: Digital Libraries and Digital Humanities

2.

In recent years the digitization of parliamentary documents increased enormously. Not only contemporary proceedings were published online, also historical debates are gradually becoming accessible through the internet. Hansard[1], for instance, made available all parliamentary proceedings, ranging from 1800 until now. For the Dutch House of Representatives, the project '*Staten Generaal Digitaal*' is currently being finalized.[2] The Belgian parliament tries to keep up with this general trend of digitization. For several years now, photographed copies could be found on the website and the project '*plenum.be*', very recently initiated further digitization of these documents and aims to disclose all the Parliamentary Debates as plain, searchable text.[3]

Digitization could serve as a new stimulus for the study parliamentary of history and discourse. The proceedings are now available for everyone everywhere. In a few tenths of a second every historian can trace his or her particular fetish. Thank you Google. But how to cope with these huge libraries of millions of words? This problem belongs to the 'Digital Humanities'. How to study parliamentary rhetoric in all its historical immensity and detect structural patterns in this unwieldy discursive mass, patterns that are both statistically and historically relevant? In this paper I gradually develop a method that will satisfy both conditions, I hope. To accomplish this goal I will apply Data Mining techniques and Machine Learning (ML) to parliamentary discourse.

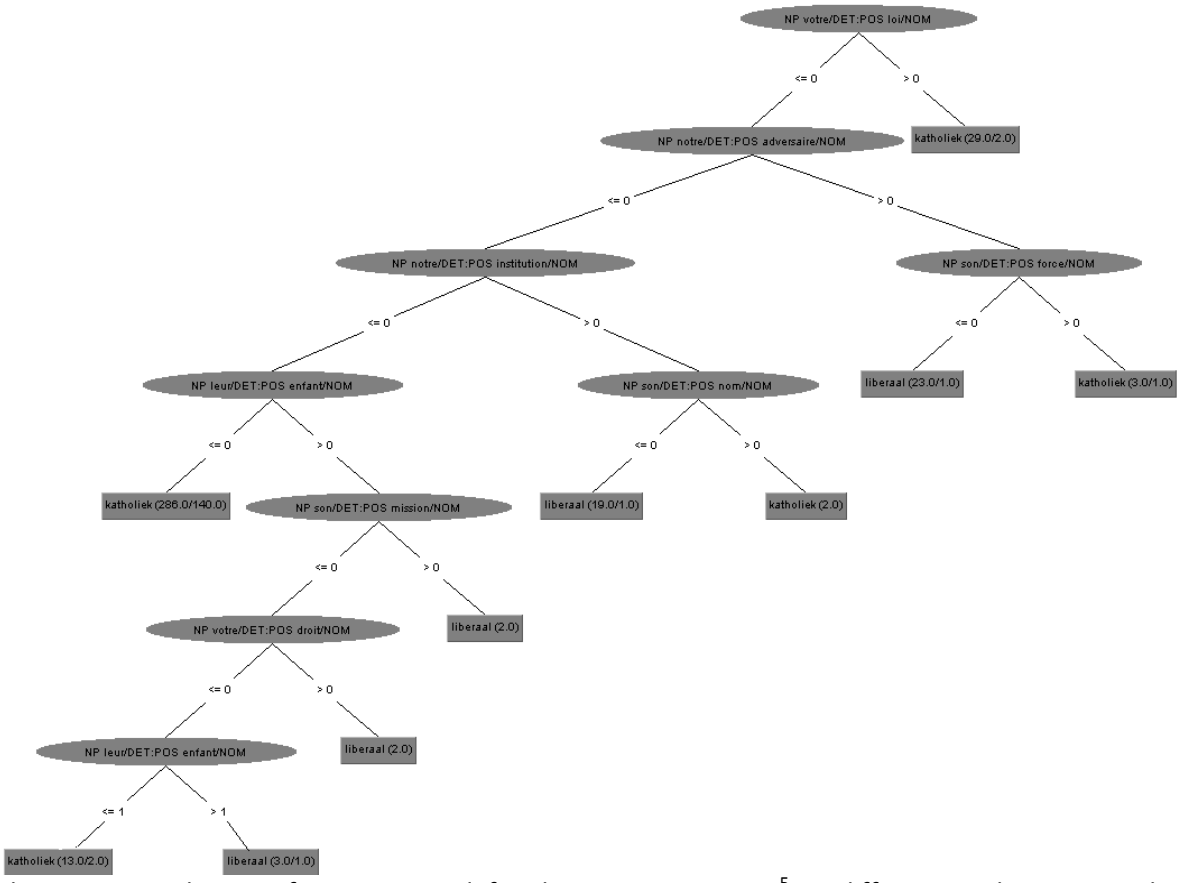
2. Theory: Machine Learning and Supervised Classification

2.1 Why Machine Learning

Why? In what respect do those methods differ from 'traditional' statistical approaches like calculating concordances, or establishing collocation tables and word frequencies? In short Data Mining is preoccupied with finding structural patterns in data by using different (Machine Learning) algorithms. An example could be the following Decision Tree:

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

Figure 1: Decision Tree (J4.8); grammar = {<DET:POS><ADJ>*<NOM|NAM>+<ADJ>*}[4]



This tree uses phrases of a certain predefined syntactic structure[5], to differentiate between Catholic and Liberal discourses. To find such a structural pattern in the data a C4.5 Decision Tree algorithm was used.[6] The tree itself consists of structured 'if… then' tests and has to be read top-down, because information gain decreases with every node. So if (=branches) a text contains the expression '*votre loi*'(=node) it covers 29 Catholic speeches and 2 Liberal (=leaves) etcetera…

Argamon and Olsen formulate the distinction with other statistical models as follows '[these models] place the onus on the user to construct queries and assimilate results, without leveraging the capacity of machines to identify patterns in massive amount of data.'[7] A machine's ability to proceed data on a very large scale while scrutinizing it for patterns, is one advantage. Beside generating a tree, the algorithms calculates how accurate this pattern is for predicting if a text is 'Liberal' or 'Catholic' (how this is done will be explained later). Here exists another benefit of machine learning, according to Goulain et al, namely the 'predictive model of testable accuracy'.[8] Through training a machine acquires a certain kind of 'artificial intelligence' which enables him to recognize different kinds of texts and consecutively classify unseen documents according to a prespecified label. This

[4] 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: <http://nltk.googlecode.com/svn/trunk/doc/api/nltk.classify-module.html>. For pre-processing Python was used.

[5] Namely combinations of possessive pronouns and nouns.

[6] Witten, J & Frank, E., '*Data Mining : Practical Machine Learning Tools and Techniques*', Morgan Kauffman Publishers, 2005, p. 97 – 105.

[7] Argomon, S. & Olsen M., 'Words, Patterns and Documents: Experiments in Machine Learning and Text Analysis', *Digital Humanities Quarterly*, 3:2, 2009, p. 2.

[8] Goulain, J-B, Argomon, S. et.al., 'Vive la Différence! Text Mining Gender Difference in French Literature '*Digital Humanities Quarterly*, 3:2, 2009, p. 1.

kind of classification is known as 'supervised classification'(SC).[9] Of course, machine learning could be used for other tasks, and there exist more types of classification, but in this paper I will restrict myself to supervised classification.
The accuracy measures how efficient a machine is in performing the classifications tasks. Moreover the learner constructs a model containing features, mostly pattern or words, which are useful in discriminating texts. The features could be compared to the nodes of the foregoing tree. I will refer to them later as 'most informative' or 'most distinctive features'.

2.2 Ideology and Deixis

Furthermore machine learning lends itself better for analyzing the relationship between text and contextual elements, like gender, race, age and, as this paper will show, political ideology.[10] The reason for applying machine learning to political discourse, came from a specific question which I was asking myself during doing research on parliamentary rhetoric: 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.[11] Do rightwing MPs speak the same language as those representatives occupying the benches on the left? If true, where to place to place the linguistic divide? How wide is this divide? The idea might not be as weird as it sounds. For a long period machine learning was principally used for authorship attribution, like distinguishing Shakespeare's plays from Marlowe's.[12] Besides comparing individuals, some studies investigated collectivities, and differentiated between male or female authors, age groups or ethnicity. [13] Here I opted for analyzing discursive differences between political ideologies.
Ideology, as one of the most elusive concepts in de human sciences[14], is of course hard to quantify. In this paper I'll spend much of my attention to person deixis. Person deixis comprises personal and possessive pronouns, terms referring to contextual elements, and thereby 'anchoring' the speaker in an interpersonal social network.[15] This 'anchoring', makes the pronominal system so crucial in studying the relationship between language and ideology: 'Ideologies mark group relations and interests and, therefore, can be studied by looking at personal pronouns.'[16] Maitland Wilson posited that ideological affiliated discourses will use the same patterns of personal pronouns.[17] Beside the pronominal system, Ideologies impose other discursive limitations. Words are not freely combined but are subject to 'collocational restraints'.[18] Because of this coupling with 'habitual', 'common-

[9] See the NLTK handbook, chapter 6: <http://nltk.googlecode.com/svn/trunk/doc/book/ch06.html>
[10] See the spring edition of '*Digital Humanities Quarterly*' which is fully dedicated to Data Mining.
[11] Verschueren, J. '*Engaging with Language Use and Ideology*', Manuscript.
[12] For Example: 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.
[13] 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.
[14] Blommaert J. & Verschueren, J. 'Debating Diversity: Analyzing the Discourse of Tolerance', Routledge, 1998, p. (?).
[15] Verschueren, J. '*Understanding Pragmatics*', Oxford University Press,, 1999, p. 20; Chilton, P. 'Analyzing Political Discourse: Theory and Practice', Routledge, 2003, p. (?).
[16] Ratia, M., 'Personal Pronouns in Argumentation: An early tobacco controversy' in: Janne Skaffari (ed.), '*Opening Windows on Texts and Discourses of the Past*', Benjamins, 2005, p. 124. See: Van Dijk, T., 'Ideological Discourse Analysis', in: Ventola. E. & Solin, A., *Interdisciplinary Approaches to Discourse Analysis*, Helsinki University Printing House, 1995.
[17] 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.
[18] 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.

sensual' and 'restrictive', I try to 'catch' the ideology of a text by seeking for recurrent pronominal and syntactic patterns in language use.
Some works of Prost, Guilhaumou and Robin were another source of inspiration.[19] They applied a more rigid linguistic method to a historical analysis of political speech. Many articles bearing the same rigorous approach but focusing on different aspects of language use were published in the journal '*Mots*'. For example: A whole edition was dedicated to deixis, especially the first plural, were Annie Geoffrey studied 'we' in Robespierre's discourse and Robert Benoit analyzed the use of the same pronoun in pamphlets of the French Communist Party [20]. More recently in collaboration with '*Lexicometrica*' special attention was given to different forms of automatic textual analysis.[21]
First and foremost I was inspired by Prost's '*Vocabulaires des proclamations électorales*'. In this work he principally studied word frequencies in French '*professions de foi électorales*' at the end of the nineteenth century. By studying the vocabulary he tried to expose differences between left- and rightwing discourse. The problem Itself fascinated me, but since this work dates from the seventies, I thought there must be more advanced methods to study political discourse.
A fundamental problem with Prost and many lexicometrical articles is that they principally concentrate on the word itself, but give little or no attention to context or combinations between lexemes. The method I'll propose below, allows to infinitely vary between different patterns, character, words, noun phrases or sentences. Moreover, whether the distinction between left and right is the most succinct, remains an open question. Perhaps the vocabularies diverge more when comparing generations? And how to calculate these differences? Although my interest here is, as with Prost[22], strictly historical, the method is an application of artificial intelligence to historical documents, because statistical differences in word use do not prevail, but the extent to which a computer is able to recognize certain properties of documents based on different syntactical patterns.

2.3. Natural Language Processing and Machine Learning

The models I'll develop below, fit within the framework of 'Natural Language Processing'[23], 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.[24] Since I don't

[19] 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.
[20] Benoît, R., 'Les figures du parti : formation et définition du groupe (1932 – 1946)', *Mots*, 10, 1985, pp. 109 – 132.
[21] Geffroy, A., 'Le nous de Robespierre ou le territoire impossible', *Mots*, 10, 1985, pp. 63 – 90.

[22] 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.
[23] 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.
[24] 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.

consider myself a specialist, this paper is also written for an audience without any foreknowledge of Machine Learning or Artificial Intelligence.
Furthermore, I was inspired by artificial intelligence in the sense that I tried to turn brain-less machine into an actor who is able to take decisions independently. The acquisition of intelligence requires a learning process. But what does it exactly mean to claim that a machine can learn? The more technical definition of 'machine learning' is:

> 'Things learn when they change their behavior in a way that makes them perform better in the future.'[25]

This definition connects 'learning' to performing or 'acting'. 'Acting' here means classifying texts, which in practice amounts to assigning labels to documents. To explain how this works, I will briefly explain some subject-specific jargon, namely 'instances', 'attributes' and 'classes'?
'Instances' reduce any document to a certain number of 'attributes' or 'features'. Which features carry away our interest? My selection was based on two criteria: context and structure. The quantification of texts is always accompanied by a degree of decontextualization. By studying words in a discursive vacuum, it is impossible to reconstruct the way they were originally used. I tried to circumvent this problem to some extent by focusing on textual patterns. The simplest examples are the '*n*-grams'. These are sequences of *n* characters or words, so each term carries some contextual references.[26]
Nevertheless, only scrutinizing n-grams could be criticized as a superficial approach to discourse. To further penetrate the structure of parliamentary language, the analysis is placed on a more syntactic level, which means I'll search for regularities in the grammatical structure. Not only words itself, but also the syntactic links between word classes become paramount. Attributes then are 'grammars' or clusters of word classes, like couples of adjectives and nouns. At this stage of my research I explored some very simple grammars, but nothing will impede further expansion and refinement in the future.
Each 'instance' resorts under a certain 'class'. Classes are equivalent to the labels given to a text. This lecture devotes its attention mainly to the ideological background of the MP. For the Belgian case the overriding conflict between the 'Liberal' and 'Catholic' classes is paramount. Besides this dichotomy I will distinguish between age groups and between periods. Each of these classes will be discussed below.
This summary doesn't answer the question how a machine acquires 'intelligence' or how it is able to 'learn'. Learning is divided into two phases. First, the database is split into a 'train set' and a 'test set'. The first set, the train set, allows the classifier to calculate how, for each instance, the attributes and the classes relate to each other. Based on the knowledge gained here, it calculates the classes of unlabeled texts in the test collection . This is done using the Naive Bayes algorithm[27], whose mathematical expression is as follows.

---

[25] Witten, J & Frank, E., '*Data Mining* , p. 8.
[26] *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*'.
[27] Witten, J & Frank, E., '*Data Mining* , pp. 94 - 97. See also : Qu H., La Pietra A., Poon S., '*Automated Blog Classification: Challenges and Pitfalls*', in: *Proceedings of the Twenty-first National Conference on Artificial Intelligence*, AAAI Press, 2006. < https://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-037.pdf>, Turkel, W., '*A Naive Bayesion in the Old Bailey, part 1*' < http://digitalhistoryhacks.blogspot.com/2008/05/naive-bayesian-in-old-bailey-part-1.html>, McCallum, A. &

$$P(label|features) = \frac{P(label) x\, P(features|label)}{P(features)}$$

P (label | attributes) calculates the 'conditional probability' or the probability that a instance resorts under a particular class given the attributes. 'Naive' here refers not to stupidity but to the assumption that there is no correlation between the attributes, so a part of the formula can be rewritten as:

$$P(features|labels) = P(feature_1|label) x \ldots x\, P(feature_n|label)$$

In some cases, I opted for Multinominal Naive Bayes because the algorithm is more sensitive to word frequencies.[28] The equation takes the following form:

$$Pr[E|H] \propto N \times \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!}$$

$n_1, n_2, \ldots n_k$ is the number of times the word i occurs in the document and $N = n_1 + n_2, \ldots + n_k$. $P_1, P_2, \ldots, P_k$ is the probability of obtaining word i when sampling from all the documents in category H .[29]

By comparing the results derived from the classification algorithm to the original labels, the learner calculates how successful - or rather how accurate - it is in recognizing the classes. In other words, accuracy equals the ratio of correctly labeled files in the test set to the total number of files in the test set.[30] In what follows we are interested not only in the degree of accuracy but also in the distinctive features which enable the learner to distinguish between various classes. These features are also known as most informative or discriminative features.

3. Practice: Adversative Rhetoric and Deixis

3.1 Introduction

For now, this methodology will be tested on a relatively small corpus of around half a million words. The text material is drawn from two key parliamentary debates in Belgian political history. The first debates date from 1879, when the laicization of primary education formed the bone of contention between the parliamentary left and right. The primary goal of the liberal government, led by Frère-Orban and Van Humbeeck, was to loosen the clerical grip on primary education.[31] During this period, a predominantly philosophical divide between left-wing Liberals and right-wing Catholics cuts through the Belgian parliament.[32] Most of the calculations below are based on these debates.

During the second series of debates, those on electoral reform in 1893, the relative strength between the parties changed. Liberals were driven into minority for good after all the agitation that accompanied the 'School war'. From 1884 until the First World War, Catholics occupied the majority benches.[33] In 1893 the extreme left enforced a constitutional reform, demanding the introduction of universal suffrage. The government headed by Beernaert, found a difficult two-thirds majority in the

---

Nigam K. 'A Comparison of Event Models for Naive Bayes Text Classification', In: '*AAAI/ICML-98 Workshop on Learning for Text Categorization*', AAAI Press,. 1998 pp. 41–48.

[28] Witten, J & Frank, E., '*Data Mining* , pp. 94 – 95.

[29] Witten, J & Frank, E., '*Data Mining* , pp. 95.

[30] <http://nltk.googlecode.com/svn/trunk/doc/book/ch06.html>

[31] Witte E, Craeybeckx, J., Meynen, A., '*Politieke Geschiedenis van België: van 1830 tot heden*', Standaard Uitgeverij, 1997, p (?).

[32] Stengers, J. & Gubin Eliane, '*Le grand siècle de la nationalité belge, de 1830 à 1918*', pp. 75 – 86.

[33] See also: Lamberts, E., (ed.) '1884: un tournant politique en Belgique Colloque: Bruxelles, 24.11.1984', in : '*Travaux et recherches / Facultés universitaires Saint-Louis*', vol. 7, 1984.

universal plural vote system, hereby circumventing the doctrinal Liberals and Catholics on the extreme right.[34] The emergence of the social question further complicated the traditional philosophical antagonism between left and right.
Table 1 shows a fine and classical example of supervised classification and information extraction. To distinguish the main topics of parliamentary debates we extracted the 1000 most frequent simple noun phrases containing one noun and at least one adjective. The grammar thus receives the following form:

grammar ={(<ADJ>+<NOM|NAM><ADJ>*)}
{(<ADJ>*<NOM|NAM><ADJ>+)}

Comparing debates on primary education (1879) with those on universal suffrage (1893), leads to the following results:

Table 1: grammar = {<ADJ>*<NOM|NAM><ADJ>*}[35]

accuracy = 0,97	
expression	**likelihood ratio**[36]
suffrage/NOM **universel**/ADJ = True	**1893** : 1879 = 44.8 : 1.0
classe/NOM **ouvrier**/ADJ = True	**1893** : 1879 = 16.2 : 1.0
loi/NOM **électoral**/ADJ = True	**1893** : 1879 = 13.1 : 1.0
corps/NOM **électoral**/ADJ = True	**1893** : 1879 = 9.5 : 1.0
instruction/NOM **publique**/ADJ = True	**1879** : 1893 = 7.7 : 1.0
école/NOM **primaire**/ADJ = True	**1879** : 1893 = 7.1 : 1.0
enseignement/NOM **primaire**/ADJ = True	**1879** : 1893 = 7.1 : 1.0
classe/NOM **inférieur**/ADJ = True	**1893** : 1879 = 6.9 : 1.0
régime/NOM **parlementaire**/ADJ = True	**1893** : 1879 = 5.4 : 1.0
autre/ADJ système/NOM = True	**1893** : 1879 = 5.4 : 1.0

This table shows how syntax-based classification pins down the discussion topics quite accurately (accuracy = 97%). As expected the 1879 debates center on primary education ('*école primaire*' and '*enseignement primaire*'), while more than a decade later, the extension of voting rights ('*suffrage universel*', '*loi électoral*') and the working classes ('*classe ouvrière*', '*classe inférieure*') enter the discussion. Similar techniques could be used for classifying all parliamentary debates according to their discussion topic, and it has many applications for making accessible corpuses containing million of words.

3.2Ideology

3.2.1 *n*-Grams

But here we have other preoccupations: tracing discursive divergences between collectivities with different political preferences. Because of the rather small corpus, the results shown here a provisional. Theoretically the source material could include every word uttered in Belgian parliament. Beside my laptop's 'horsepower' there is no restriction on the size of the data. All of the following results are drawn from the debates on the reform of primary education (1879).
What are these results? Let's first examine the *n*-grams, i.e. sequences of *n* words, like the following bigrams: [enough theory],[theory, what],[what, are]…

[34] Van Eeno, R., 'Kiesstelsels en verkiezingen, 1830-1914', in: Gerard, E., Witte, E., Gubin, E., & Nandrin, J. P. (eds.), *Geschiedenis van de Belgische Kamer van Volksvertegenwoordigers 1830-2002*, , Kamer van Volksvertegenwoordigers, 2003
[35] Classifier = Naive Bayes, validation = 50 % train set, 50 % test set
[36] The left side of the table has to be read as follows : 'kathol : libera = 12.9 : 1.0' means the expression is 12,9 times more likely to appear (the 'True' condition) in a Catholic then in a Liberal document.

Table 2:accuracy for *n*-Gram-based classification (1879)[37]

Length *n*-gram	Accuracy
1	**0.78**
2	**0.78**
3	**0.80**
4	0.75
5	0.60
6	0.57
7	0.54
8	0.51
9	0.52

The accuracy doesn't decrease as the length of ngram increases, as might be expected. Trigrams seem to deliver the best results with an average accuracy of 80%. Afterwards the accuracy declines gradually.If we take a closer look at the trigrams we get the following results:

Table 3[38]: Trigrams (1879 selection)

trigram	likelihood ratio
l honorable rapporteur = True	Catholic : Liberal = 12.9 : 1.0
au nom de = True	Catholic : Liberal= 9.5 : 1.0
m le ministre = True	Catholic : Liberal= 7.7 : 1.0
de la gauche = True	Catholic : Liberal= 6.7 : 1.0
de la droite = True	Liberal : Catholic = 6.0 : 1.0
de la majorité = True	Catholic : Liberal= 5.5 : 1.0
de la chambre = True	Catholic : Liberal= 5.5 : 1.0
de la liberté = True	Catholic : Liberal= 5.1 : 1.0
de nos adversaires = True	Liberal : Catholic = 4.9 : 1.0

Constructions consisting of '*de la*/*le*' combined with a noun are very effective in categorizing the data. The debates of 1879 are permeated with an ever present opposition between 'us' and 'you', a divide that runs parallel to the gap between the two parties. This bipolar and 'adversative' nature is reflected in constantly talking about the Other. '*Monsieur le ministre*' '*de la gauche* 'and '*de la majorite*' typifies the discourse of the right-wing opposition, while Liberals principally refer to '*de nos adversaires*' and '*de la droite*'.[39] This might seem a rather trivial result , but is it? The classifier consistently points to the importance of adversative language use in parliamentary discourse. A similar result was obtained by studying nouns and names. (see below) Speeches were best differentiated by the more formal aspects of parliamentary rhetoric, such as titles. Catholics focus on the liberal Cabinet ('*honorable ministre*') while their political opponents indirectly address their opponents with '*honorable monsieur*' or '*honorable membre'*.

[37] This result is obtained from a ten times repeated classification task. The 1000 most frequent *n*-Grams were chosen as features. Before Classification the documents were randomized every time and split into a 50% train set and 50% test set.

[38] Idem footnote 29

[39] 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.

3.2.2. Exploring grammars

Table 4: accuracy per word class (1879)

	<ADJ>		<NOM|NAM>		<VER:.+>	
classifier	**word**	**lemma**	**word**	**lemma**	**word**	**lemma**
MultiNominalNaiveBayes	72,77	72,51	82,98	**83,77**	**64,66**	**74,87**
J48	65,18	59,94	69,10	71,47	60,20	61,78
***k*-NN (*k*=16)**	54,71	53,40	63,61	50,50	51,57	50,52

This table shows how the Multinominal Naive Bayes scores consistently higher than the J48 Decision Tree or the k-Nearest Neighbor algorithm[40]. More important at this point is the distinction between 'words' and 'lemmas'. Lemmas[41] could be compared to entries of dictionaries, because they reduce words to their basic form. Lemmatizing boosts accuracy especially when it comes to classifying verbs. Adjectives tend to be an exception though, a phenomenon for which I currently have no decent explanation. Still I'll mostly use the lemmas for the classification tasks.
The assertion concerning adversative discourse could find further support when looking at clusters consisting of nouns or proper names and combinations of nouns and adjectives.

Table 5 : grammar = {<NOM|NAM>} (1879)[42]

Accuracy =76,44			
Liberal		**Catholic**	
x_i	$f(x_i)$	x_i	$f(x_i)$
clergé/NOM	9,28	***école***/NOM	7,79
église/NOM	3,88	***moralité***/NOM	4,74
autorité/NOM	3,48	***liberté***/NOM	4,24
malou/NAM	2,99	***commune***/NOM	4,21
prêtre/NOM	2,85	***état***/NOM	3,01
évêque/NOM	2,41	***enfant***/NOM	2,85
opposition/NOM	1,78	***olin***/NAM	2,71
loi/NOM	1,66	***éducation***/NOM	2,29
woeste/NAM	1,53	***loge***/NOM	1,90
		rapporteur/NOM	1,86
		constitution/NOM	1,83

The proper names 'Malou', 'Woeste' and 'Beernaert' all rightwing MPs, are more likely to appear in liberal speeches, while representatives on the right more often drop names of their left wing adversaries 'Van Humbeeck' and 'Olin', or they refer to functions like '*ministre*' or '*rapporteur*', that at the time were met by liberals. The same is true for other nouns. The left half 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 other nouns like '*église*', '*prêtre*' and '*évêque*'. On the other half of the hemisphere, the State ('*État*'), the free masons ('*loge*') and the liberal majority ('*majorité*')[43] have to suffer for their sins. The discourses on both

[40] Notoriously missing here is a Suport Vector Machines, which is often used for classification tasks. I'm aware of this defect, but due to technical problems en intellectual deficiencies I wasn't able to include SVM in my research for the moment

[41] TreeTagger was used for lemmatization : See <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/> It is developed by Helmut Schmidt at the University of Stuttgart.

[42] Multinominal Naive Bayes and 10 fold cross-validation was used for calculating the accuracy. Features are 1000 most frequent patterns as defined by the grammar.

[43] Grammars consist of word classes, (which are annotated as follows 'NOM'=noun; 'NAM'=name; 'ADJ'='adjective'; 'VER'=verb; ,'PRO:PER'=personal pronoun; 'DET:POS'=possessive pronoun) and operators

sides differ in the characterization of the Other. *Hétérodésignation* rather then *autodésignation* defines this kind parliamentary discourse.[44]

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 the guardians of the municipal powers. They defend municipalities against a State who's growing power threatens to monopolize society ('*monopole*').According to them, the devolution of authority is sanctioned by the Constitutional Congress ('*congrès*', '*constitution*'). Belgium inherited a specific character ('*caractère*') at its birth, a constitutional design that is being menaced by liberal modernist delusions.

Furthermore, there are some small lexical differences, such as the Liberal preference for '*foi*' instead of '*réligion*', or more symptomatic in this context, is the Liberal choice for '*enseignement*' were Catholics opt for terms like '*école*', '*education*', or '*enfant*'. The last two terms reach very high $f(x_i)$ values. Maybe the Catholic support for the delegation of power to the lower institutional *echolons*, is connected with the a predilection for concrete elements as 'the child' or 'the school'. (see below)

Table 6: grammar ={(<ADJ>+<NOM|NAM><ADJ>*)}
{(<ADJ>*<NOM|NAM><ADJ>+)} (1879)[45]

accuracy = 74,53

Liberaal >		**Katholiek >**	
x_i	$f(x_i)$	x_i	$f(x_i)$
honorable/ADJ ***monsieur***/NOM	23,45	***honorable***/ADJ ***ministre***/NOM	19,88
honorable/ADJ ***membre***/NOM	12,59	***instruction***/NOM ***publique***/ADJ	16,10
autorité/NOM ***civil***/ADJ	4,55	***section***/NOM ***central***/ADJ	9,99
parti/NOM ***libéral***/ADJ	3,55	***comité***/NOM ***scolaire***/ADJ	8,95
opinion/NOM ***libéral***/ADJ	3,14	***école***/NOM ***libre***/ADJ	5,30
pouvoir/NOM ***public***/ADJ	2,80	***morale***/NOM ***chrétien***/ADJ	3,82
croyance/NOM ***religieux***/ADJ	2,32	***morale***/NOM ***universel***/ADJ	3,78
parti/NOM ***catholique***/ADJ	2,15	***défense***/NOM ***national***/ADJ	2,70
clergé/NOM ***catholique***/ADJ	1,89	***état***/NOM ***libéral***/ADJ	2,04
		système/NOM ***anglais***/ADJ	1,78
		morale/NOM ***moyen***/ADJ	1,74
		école/NOM ***catholique***/ADJ	1,70

From a cursory analysis of the distinctive elements emerges the following picture. 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 only 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 because it is always has to be rooted in religion or tradition. The abstract ethic which the left promotes, tend 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 its will on a majority of the Catholics. In

('|'=or; '+'=one or more occurrences of the previous entity; '*'= zero or more occurrences of the previous entity).

[44] Geffroy, A. 'Le *nous* chez Robespierre ou le territoire d'impossible', *Mots*, 10, 1985, p. 63.

[45] Validation method: 10-fold cross-validation, Classifier: Multinominal Naïve Bayes

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'Etat*') and '*école libre*', which coincides with '*école catholique*'.

Table 7 : Confusion Matrix,grammar = {<NOM|NAM>} (1879)

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

Table 8: Confusion Matrix, grammar = {<ADJ>} (1879)

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

Table 9 : 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[46] show how the classifier consistently has more difficulties with recognizing Liberal than Catholic discourse. It classifies more Liberal speeches as Catholic than the other way around. The Catholic discourse seems to be more coherent. Historically, 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..

3.2.3 Exploring Person Deixis

The conflict between 'us' and 'you' that underlies the debates, leads to a close study of person deixis. To point to differences in identification and alterization I investigated the use of the possessive pronouns '*notre*' and '*votre*' and it yielded the following results:

Table 10 : grammar = {<NOTRE><NOM>} grammar = {<VOTRE><NOM>} (selection) (1879)[47]

av.accuracy = 0.52		**av.accuracy = 0.57**	
expression	**likelihood ratio**	**expression**	**likelihood ratio**
nos adversaires = True	Liberal : Catholic = 4.7 : 1.0	**votre loi** = True	Catholic : Liberal= 10.3 : 1.0
nos enfants = True	Catholic : Liberal= 2.8 : 1.0	vos instituteurs = True	Catholic : Liberal= 2.3 : 1.0
nos populations = True	Catholic : Liberal= 2.8 : 1.0	votre système = True	Liberal : Catholic = 2.3 : 1.0
nos institutions = True	Liberal : Catholic = 2.6 : 1.0	**votre parti** = True	Liberal : Catholic = 2.3 : 1.0
notre pays = True	Catholic : Liberal= 2.5 : 1.0	**vos amis** = True	Liberal : Catholic = 1.8 : 1.0
nos croyances = True	Catholic : Liberal= 2.2 : 1.0	vos mains = True	Liberal : Catholic = 1.7 : 1.0
nos écoles = True	Liberal : Catholic = 2.0 : 1.0	vos enfants = True	Liberal : Catholic = 1.7 : 1.0
nos principes = True	Liberal : Catholic = 1.8 : 1.0	votre droit = True	Liberal : Catholic = 1.7 : 1.0
nos lois = True	Liberal : Catholic = 1.8 : 1.0	**vos rangs** = True	Catholic : Liberal= 1.7 : 1.0
nos ancêtres = True	Catholic : Liberal= 1.8 : 1.0		

What is 'ours'? Liberals focus on institutions and principles, while the right side 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. In other places 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*',

[46] Results are confirmed by a classification using each word class separately.
[47] Validation method: 10 times repeated randomization, Classifier: Naïve Bayes

'*nos populations*' and '*notre pays*' Not so much the institutions are central as well more concrete communities.
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 cannot distance themselves far enough from . 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.
To get a more exhaustive picture of personal pronouns, I included verb forms using the following rather simple grammar:

grammar = (<PRO :PER> <VER.+> +)

The output looks like this:

Table 11: grammar = (<PRO :PER> <VER.+> +) (selection) (1879)[48]

Liberaal >		**Katholiek >**	
accuracy = 59,81 (MNB)			
x_i	$f(x_i)$	x_i	$f(x_i)$
nous/PRO:PER **avoir**/VER:pres	8,60	**vous**/PRO:PER **avoir**/VER:futu	3,64
je/PRO:PER **parler**/VER:pres	3,19	**on**/PRO:PER **vouloir**/VER:pres	3,44
je/PRO:PER **croire**/VER:pres	2,79	**je**/PRO:PER **demander**/VER:pres	2,76
vous/PRO:PER **avoir**/VER:impf	2,42	**je**/PRO:PER **comprendre**/VER:pres	2,30
vous/PRO:PER **être**/VER:pres	2,32	**je**/PRO:PER **avoir**/VER:pres	2,28
je/PRO:PER **faire**/VER:pres	2,20	**nous**/PRO:PER **dire**/VER:pres	1,84
on/PRO:PER **pouvoir**/VER:pres	1,92	**je**/PRO:PER **dire**/VER:impf	1,83
il/PRO:PER **dire**/VER:futu	1,77	**vous**/PRO:PER **devoir**/VER:pres	1,82
je/PRO:PER **vouloir**/VER:cond	1,54	**nous**/PRO:PER **dire**/VER:infi	1,61
nous/PRO:PER **occuper**/VER:pres	1,53	**on**/PRO:PER **nous**/PRO:PER dire/VER:pres	1,59
vous/PRO:PER **avoir**/VER:pres	1,48	**nous**/PRO:PER **croire**/VER:pres	1,39

Besides 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 plural verb forms, it is possible to map this these collective 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.

Table 12: verbs first and second person plural (1879)[49]

lemma	**av.acc. = 0,55 (NB)**	**lemma**	**av.acc.=0,55 (NB)**
nous devoir = True	Catholic : Liberal = 4.5 : 1.0	**vous dire = True**	Catholic : Liberal = 4.2 : 1.0
nous vouloir pas = True	Liberal : Catholic = 3.3 : 1.0	**vous entendre = True**	Liberal : Catholic = 3.0 : 1.0
nous proposer = True	Liberal : Catholic = 3.2 : 1.0	vous mettre = True	Catholic : Liberal = 3.0 : 1.0
nous croire = True	Catholic : Liberal = 2.9 : 1.0	**vous tenir = True**	Catholic : Liberal = 3.0 : 1.0
nous défendre = True	Liberal : Catholic = 2.6 : 1.0	**vous prétendre = True**	Liberal : Catholic = 2.6 : 1.0
nous demander = True	Catholic : Liberal = 2.6 : 1.0	**vous admettre = True**	Liberal : Catholic = 2.5 : 1.0
nousvouloir que = True	Catholic : Liberal = 2.3 : 1.0	vous invoquer = True	Liberal : Catholic = 2.5 : 1.0
nous croire que = True	Catholic : Liberal = 2.1 : 1.0	**vous prendre = True**	Catholic : Liberal = 2.4 : 1.0
nous pouvoir = True	Liberal : Catholic = 1.8 : 1.0	vous savoir que = True	Catholic : Liberal= 2.2 : 1.0
		vous parler = True	Liberal : Catholic = 2.2 : 1.0
		vous espérer = True	Liberal : Catholic = 1.8 : 1.0

[48] Validation method: 10-fold cross-validation, Classifier: Multinominal Naïve Bayes
[49] Validation method: 10 times repeated randomization, Classifier: Naïve Bayes

Among the verbs that Catholics often use, firstly comes '*nous devons*', followed by '*nous croyons*' and '*nous demandons*'. A possible interpretation lies in the tension between opposition and majority. The right rather 'asks' than 'proposes'('*nous proposons*') as does the Liberal Party. These demands are related to an obligation ('*nous devons*') and a 'belief' ('*nous croyons*'). The liberal discourse revolves around '*nous pouvons*' and '*nous voulons*'. Their main task is to protect ('*nous défendons*') of the Belgian institutions against the subversive activities of the clergy and their parliamentary fifth column.
The second person plural 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 supposedly the liberal tendency to continually appropriate and keep for themselves. Politics is a game of taking and retaining.

4. Beyond the Ideological Divide

Machine learning allows us to go beyond the ideological divide. Parliament is not a place where only parties clash, but where people with multiple backgrounds confront each other. On the benches are sitting different generations, elected in a rural or an urban constituency, cherishing their own personal convictions and preferences which don't always match the opinion of other party members. It's important to take into account these different identities. For now however, I limit myself to some little experiments, but – again – it's only the tip of the iceberg.

4.1. Comparing Generations

Until now the ideological contrast between left and right prevailed, a somehow arbitrary choice of course. Perhaps the biggest discursive differences exist between the various parliamentary generations and not between the parties? For every speech the date of birth of the MP was added to the corpus. To obtain a uniform distribution the database was split into those born before 1827 and all those and those born thereafter. A quick test with different syntactic patterns using the Multinominal Naive Bayes algorithm gave a negative result. The classifier scored consistently lower. At present it is difficult to know whether this negative result depends on the small size of the database, or that indeed party trumps age. Still, what remains an interesting question is if the generations differ in their use of personal pronouns. Maybe older MPs, prefer to represent themselves more as independent individuals, speaking in first person singular, and younger generations prefer to speak on behalf of others, using 'we' instead. Although the a rather low accuracy of[50], doesn't allow us to be really confident of our conclusions, the use of personal pronouns among older MPs showed a preference towards first person singular ('*je*': 0,056, '*mon*': 0,029[51]) as opposed to younger representatives, who more frequently speak in the first person plural. Replacing date of birth by year of entry in Parliament, leads to a similar resultMore research over a longer period and on different kinds of debates is needed however.

4.2 Classifying Diachronically

Machine learning is not limited to scrutinizing discursive divergences between groups, it also allows to uncover structural changes over time. Classifying noun phrases containing '*notre*' results in following output:

[50] 55% using Naive Bayes or 60% using a more advanced Support Vector Machine, Validation Method = 10-fold cross-validation
[51] Attribute Weights, using Multinominal Naive Bayes

Table 13: grammar = {<NOTRE><ADJ>*<NOM>+<ADJ>*}

accuracy = 0,65	
expression	**likelihood ratio**
notre/NOTRE **adversaire**/NOM = True	**1879** : 1893 = 7.2 : 1.0
notre/NOTRE **école**/NOM = True	**1879** : 1893 = 6.2 : 1.0
notre/NOTRE **collègue**/NOM = True	**1893** : 1879 = 5.8 : 1.0
notre/NOTRE **honorable**/ADJ **collègue**/NOM = True	**1893** : 1879 = 4.4 : 1.0
notre/NOTRE **liberté**/NOM = True	**1879** : 1893 = 3.7 : 1.0
notre/NOTRE **pays**/NOM = True	**1893** : 1879 = 3.4 : 1.0
notre/NOTRE **programme**/NOM = True	**1893** : 1879 = 2.4 : 1.0
notre/NOTRE **enfant**/NOM = True	**1879** : 1893 = 2.3 : 1.0
notre/NOTRE **institution**/NOM **national**/ADJ = True	**1879** : 1893 = 2.3 : 1.0
notre/NOTRE **population**/NOM = True	**1879** : 1893 = 2.2 : 1.0
notre/NOTRE **constitution**/NOM = True	**1879** : 1893 = 1.8 : 1.0
notre/NOTRE **principe**/NOM = True	**1893** : 1879 = 1.7 : 1.0
notre/NOTRE **nationalité**/NOM = True	**1879** : 1893 = 1.6 : 1.0

The shift from 'our opponents' ('*notre adversaire*') to 'our colleagues' ('*notre* (*honorable*) *collègue*'). is the most striking trend. The emphasis in 1893 is less on conflict with the opponents 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'. 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*' occurred more often than in 1893. The constitution was for both parties to of key importance to guarantee the freedoms ('*notre liberté*'), institutions ('*notre institution national*') and national identity ('*notre nationalité*').[52] The identification with these elements was less frequent during the later discussions, where mainly the expression '*notre pays*' was used.

Table 14 : grammar = {<VOTRE><ADJ>*<NOM><ADJ>*}[53]

accuracy = 0,54	
expression	**likelihood ratio**
votre/VOTRE **loi**/NOM = True	**1879** : 1893 = 9.6 : 1.0
votre/VOTRE **projet**/NOM = True	**1879** : 1893 = 3.6 : 1.0
votre/VOTRE **système**/NOM = True	**1893** : 1879 = 2.6 : 1.0
votre/VOTRE **discours**/NOM = True	**1893** : 1879 = 2.4 : 1.0
votre/VOTRE **droit**/NOM = True	**1879** : 1893 = 2.3 : 1.0
votre/VOTRE **part**/NOM = True	**1893** : 1879 = 1.8 : 1.0
votre/VOTRE **principe**/NOM = True	**1893** : 1879 = 1.7 : 1.0
votre/VOTRE **propre**/ADJ **parti**/NOM = True	**1879** : 1893 = 1.6 : 1.0
votre/VOTRE **ami**/NOM = True	**1879** : 1893 = 1.4 : 1.0

What about changes in Othering? It's hard to draw any conclusion from this table. The accuracy scores poorly since the algorithm has less material to calculate. Startling is that the syntactic combinations with the lemma '*votre*' decrease with approximately 40%, while expressions containing '*nos*' or '*notre*' increase some 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 debates French

[52] See : Stengers, J. & Gubin Eliane, '*Le grand siècle de la nationalité belge, de 1830 à 1918*', pp 8 – 14.
[53] Evaluation method = 50% test set ; 50% train set

debates on compulsory primary education (1880 – ’81) with those on educational reform in Belgium (1879).

Table 16: grammar = {(<ADJ>+<NOM|NAM><ADJ>*)}
{(<ADJ>*<NOM|NAM><ADJ>+)}[54]

accuracy = 87,18

France >		**Belgium >**	
x_i	$f(x_i)$	x_i	$f(x_i)$
NP instruction/NOM **primaire/**ADJ	23,77	**NP conseil/**NOM **communal/**ADJ	13,46
NP instruction/NOM **religieux/**ADJ	17,85	**NP section/**NOM **central/**ADJ	10,09
NP contrainte/NOM **légal/**ADJ	7,72	**NP école/**NOM **communal/**ADJ	7,83
NP enseignement/NOM **religieux/**ADJ	7,01	**NP école/**NOM **normal/**ADJ	5,45
NP école/NOM **publique/**ADJ	6,51	**NP comité/**NOM **scolaire/**ADJ	4,67
NP ancien/ADJ **régime/**NOM	4,12	**NP défense/**NOM **national/**ADJ	4,44
NP révolution/NOM **français/**ADJ	3,75	**NP parti/**NOM **libéral/**ADJ	3,54
		NP administration/NOM **communal/**ADJ	2,94

Notwithstanding the small database, it’s possible to distinguish different 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 lead to satisfactory 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.

5. Conclusion

At beginning of this paper I proposed two reasons for applying Machine Learning to parliamentary discourse. One was the availability of immense digital libraries and the ability of machines to find structural patterns in these huge datasets. A second reason was 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 adversative character of parliamentary rhetoric, although some indicators marked a shift towards promoting ingroup coherence in the later period.

But everything written here is merely the tip of the iceberg. This methodology asks for further improvement. Especially a lot of work has to be done on information extraction, on relating entities in sentences. Syntactical classification here forms a good starting point, but still I’m still far removed from the endpoint, if there is one. Furthermore, I want to experiment more with advanced classification algorithms like Support Vector Machines, although there exists a unofficial rule in data mining which claims that more complex algorithms don’t necessarily deliver better results.[55] Still, this asks for further confirmation.

[54] Evaluation method = 10-fold cross-validation; Classifier = MultiNominal Naïve Bayes

[55] Witten, J & Frank, E., ‘*Data Mining* , pp