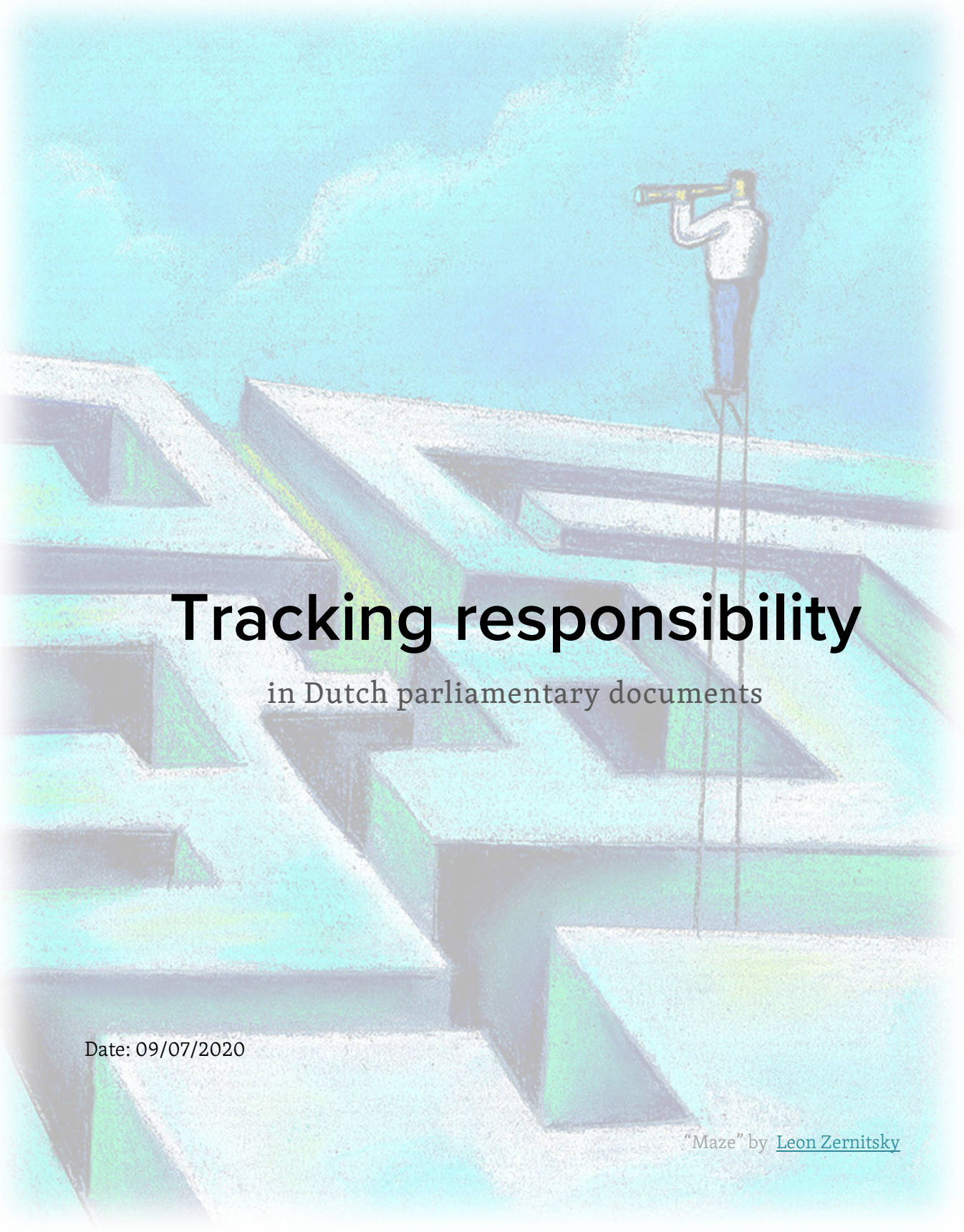
Inge Wertwijn, 13102702



Tracking responsibility

in Dutch parliamentary documents

Date: 09/07/2020

"Maze" by Leon Zernitsky

CONTENTS

LIST OF FIGURES

Abstract

This paper reports the results of a pilot study into authorship and truthfulness of parliamentary documents, using stylometric methods for language analysis.

In a corpus of 40 related documents from the same parliamentary dossier, authorship can probably be attributed to at least 27 (groups of) authors. In one case, one and the same author appears to have written closely related documents both on behalf of the (prime)minister and of the General Audit chamber. If this finding were to be established beyond reasonable doubt, this would be severely frowned upon.

It was not possible to connect established measures for deceptive language to parliamentary documents. If anything, parliamentary appear to be more trustworthy than the control set, even when we know for certain that some documents have deceptive contents. It appears that there is quite a large general style difference between parliamentary documents and more texts that should be considered before any further investigation.

1 Research question

1.1 Authorship

Dutch parliamentary documents are usually written on behalf of a minister, state secretary, committee, ministry, or office, i.e. the party or function formally responsible. The name of the person who wrote the text is rarely recorded on official documents. It is therefore not known what or how many people are involved in writing parliamentary documents. This project attempts to shed some light on this.

1.2 Truthfulness

It is known that parliamentary documents sometimes contain half-truths, falsehoods, or omissions. It is not known whether deception can be detected in parliamentary documents. This project investigates whether some well-known markers for deception can be detected in parliamentary texts.

2 Corpus Preparation

2.1 Primary selection

Since parliamentary documents are public, there are no copyright or availability issues. All parliamentary documents addressed to the Dutch House of Representatives are accessible via www.overheid.nl.

For this pilot study, a selection was made of 53 documents, all belonging to dossier 31066 (Dutch Tax Office), published from 2017 up to 2020 on behalf of various parties. This dossier contains quite a few documents which in hindsight proved to be not entirely truthful. These documents are part of the so called "Toeslagen affaire", parents persecuted for alleged fraud by the Dutch Tax office

As www.overheid.nl does not offer a bulk download service, these 53 documents were downloaded by hand. The resulting set of documents contained between 417 and 88.808 words, counted using *Count Anything*, 2009.

Filename	Words	Chars	Chars no spaces
blg-893312_2019-07-23_MinFin.pdf	417	2942	2552
blg-922421_2019-12-17_Ombudsman.pdf	420	2736	2388
kst-31066-604_2020-03-02_Staatssecretarissen van Financien.pdf	436	2816	2451
kst-31066-605_2020-02-07_Staatssecretaris van Financien.pdf	538	3220	2756
kst-31066-577_2019-12-11_Staatssecretaris van Financien.pdf	574	3815	3334
kst-31066-589_2020-01-17_Minister van Financien.pdf	685	4573	3996
kst-31066-611_2020-03-12_Staatssecretaris van Fianancien.pdf	748	4793	4157
kst-31066-539_2019-11-20_Staatssecretaris van Financien.pdf	942	6038	5238
kst-31066-401_2018-04-16_Staatssecretaris.pdf	992	6305	5467
kst-31066-438_2018-10-31_Staatssecretaris.pdf	1002	6420	5587
blg-922422_2020-02-03_Belastingdienst.pdf	1155	7540	6561
kst-31066-594_2020-01-24_Minister van Financien.pdf	1191	7379	6375
kst-31066-524_2019-09-20_Staatssecretaris van Financien.pdf	1366	8799	7647
blg-922420_2020-01-30_Minister van Financiën & Minister President.pdf	1387	8572	7382
kst-31066-534_2019-10-31_Staatssecretaris van Financien.pdf	1468	9726	8453
kst-31066-533_2019-10-30_Staatssecretaris van Financien.pdf	1737	11271	9801
kst-31066-480_2019-0417_Staatssecretaris.pdf	1749	11532	10057
blg-917857_2-17-11-07_Belastingdienst.pdf	1831	11773	10046
blg-908390_2019-10-28_ADR.pdf	1994	14704	12853
kst-31066-596_2020-02-04_Minister en Staatssecretaris van Financien.pdf	2113	13875	12076
kst-31066-588_2020-01-11-Minister van Financien.pdf	2496	16368	14261
kst-31066-574_2019-12-17_Staatssecretaris van Financien.pdf	2890	19022	16597
blg-826385_2017-12-12_MinFin.pdf	3178	22775	20030
kst-31066-599_2020-02-13_ARK.pdf	3260	21771	18981
blg-925925_2019-12-13_Staatssecretaris van Financien.pdf	3337	20759	17791
kst-31066-444_2018-12-04_Staatssecretaris.pdf	3504	23068	20190
kst-35302-26_2019-11-08_Staatssecretaris van Financien.pdf	3611	23216	20228
kst-31066-330_2017-01-27_Minister en Staatssecretaris van Financien.pdf	4795	32235	28138
blg-855692_2018-09-19_Belastingdienst.pdf	5363	38318	33759
kst-31066-603_2020-03-02_Staatssecretarissen van Financien.pdf	5503	35495	30792
kst-31066-609_2020-02-27_Staatssecretarissen van Financien.pdf	5560	37402	32768
kst-31066-403_2018-04-26_Staatssecretaris van Financien.pdf	5805	39285	34360
blg-920774_2020-01-17_ABD.pdf	6029	39915	34898
kst-31066-495_2019-06-17_Staatssecretaris.pdf	6110	40204	35149

blg-893311_2019-07-23_Belastingdienst.pdf	6354	43869	38763
kst-31066-607_2020-02-27_Staatssecretarissen van Financien.pdf	6906	45089	39254
kst-31066-408_2018-06-11_Staatssecretaris.pdf	6934	44454	38687
blg-908392_2019-11-11_Belastingdienst.pdf	7705	52706	46505
blg-893312_2019-07-23_ADR.pdf	8317	56359	48985
blg-840575_2018-05-01_ADR.pdf	8501	57771	50004
kst-31066-538_2019-11-15_Staatsecretaris van Financien.pdf	9164	59856	52004
blg-884750_2019-06-03-ADR.pdf	10473	70640	61528
blg-920773_2020-01-17_ABD.pdf	12617	86446	75943
blg-826046_2017-12-07_Belastingdienst.pdf	14282	101499	88197
blg-880514_19-04-2019_Belastingdienst.pdf	15098	103378	91124
blg-805595_2017_04_24_Belastingdienst.pdf	16118	113385	100377
blg-920771_2019-12-03_KPMG.pdf	16684	375331	362122
blg-862319_2018-11-15_Belastingdienst.pdf	16688	113141	99784
blg-926527_2020-03-12_ADR.pdf	19321	122804	106491
blg-839367_2018-04-18_Belastingdienst.pdf	23687	156076	141810
blg-914023_2019-11-14_Adviescommisie Uitvoering Toeslagen.pdf	33891	224310	192437
blg-926526_2020-03-12_Adviescommisie Uitvoering Toeslagen.pdf	54788	365919	314154
blg-920464_2019_10_31_EY.pdf	88008	495124	413823

Figure 1 Corpus of parliamentary documents showing number of words and characters

Extremely large (>20.00 words) and extremely small (<1000 words) documents were left out. These are show in red in the table above. This reduced the set of documents to 40. These were converted from pdf to txt (UTF8 format) using the batch function of *Adobe Acrobat Pro DC*, 2020. There were no documents that needed OCR.

The converted documents were renamed to allow for effective visualisation, following this pattern:

Responsible-party _dossier-number#yymmdd.txt

2.2 Control set

To compare the set of parliamentary documents with a general sample of other documents, a second set of 20 texts was retrieved from the internet. This control set contained:

- articles by the Dutch magazine Follow the Money, on the topic of the Belastingdienst (freely accessible on the first day of publication at www.ftm.nl)
- Short stories by Biesheuvel, Carry van Bruggen and Schendel from www.dbnl.org (freely accessible)
- Short stories by contemporary Dutch authors, from www.hebban.nl (freely accessible)

Documents were selected to fit between 1000 and 20.000 words, the boundaries that were set for the primary set of documents. As with the primary set of documents, words were counted using *Count Anything*, 2009

Filename	Words	Chars	Chars no spaces
xx-Biesheuvel_brommeropzee.txt	1625	8690	7188
xx-schendel_broosgeluk.txt	2059	12002	10037
xx-Biesheuvel_dewereldmoetbeterworden.txt	2242	12578	10526
xx-Lisdonk_payback.txt	2424	13607	11470
xx-Appel_metgelijkemunt.txt	2481	14141	11942

xx-Coolwijk_stik.txt	2486	14596	12403
xx-Broersma_Solitairet.txt	2498	13681	11528
xx-Beek_derit.txt	2527	14084	11859
xx-Hendriks_withorwithoutyou.txt	2536	14245	11996
xx-Schendel_maneschijn.txt	2902	15975	13239
xx-Maron_eenbizarongeluk.txt	3327	19311	16344
xx-ftm_Wegkijken van commissie-Donner.txt	3633	23909	20496
xx-Bruggen_vadersboek.txt	3911	21865	18231
xx-ftm_Georganiseerde misdaad roofde miljarden.txt	4610	31372	27040
xx-ftm_Politieke top was gewaarschuwd .txt	4610	31372	27040
xx-Bruggen_eenheledonderdagthuis.txt	4888	27833	23138
xx-Bruggen_voetvandenijsberg.txt	4989	28478	23673
xx-Schendel_blidmonde.txt	6854	39544	33115
xx-Heuvelt_devisindefles.txt	10222	58454	49049
xx-Schendel_hetvertrouwen.txt	12196	69412	57862

Figure 2 Control set: corpus of random texts showing number of words and characters

Documents were named following this pattern:

xx-Author _Storyname.txt

The purpose of using the xx-suffix is to distinguish control-documents from primary documents.

2.3 WORDLIST AND CULLING

Because parliamentary documents contain a lot of jargon and (institutional) names, this may distort word frequency counts. These jargon words needed to be identified so that they can be left out from the analysis, as a form of culling.

To achieve this, AntConc[1] was used. AntConc is a freeware software tool for corpus linguistics research, containing seven different tools to reveal patterns in large scale textual objects. Using AntConc an all-lowercase wordlist was generated from the parliamentary document set containing the 5000 most frequently used words, without sampling, i.e. using the full texts. This list was loaded into *Microsoft Excel 365*, 2020. Jargon and (institutional) names were then marked and saved as a separate list of 499 words, as show below.

Jargon: aanbiedingsbrief, aangifte, aangiftebehandeling, aangiftecampagne, aangifteformulier, aangiften, aangiftenbehandeling, aangiftes, aanmaning, aanmaningskosten, aanslag, aanslagoplegging, aanslagregeling, aansprakelijkheidsbepaling, aansprakelijkstelling, accijns, accijnzen, adviescommissie, afdrachtsvermindering, afdrachtsverminderingen, aftrekposten, afvalstoffenbelasting, afwijkingsgrond, aix, amlc, anbi, anpr, apbi, art, assurantiebelasting, atle, autobelastingen, autobrief, autodomein, autoheffingen, automiddelen, avg, awb, back, baliebrief, bdate, begroting, begrotingsartikel, behandelbundel, behandelproces, behandelteams, belastingaangifte, belastingaanslagen, belastingafspraken, belastingbedrag, belastingcontrole, belastingdienstbrede, belastingdienstbreed, belastinggrondslag, belastingheffing, belastinginkomsten, belastingjaar, belastingmiddel, belastingmiddelen, belastingontvangsten, belastingopbrengsten, belastingplan, belastingplanpakket, belastingplantraject, belastingregeling, belastingrente, belastingschuldigen, belastingsubject, belastingvorderingen, belastingwetten, belastingzaken, bes, beslagopdrachten, beslagvrije, betalingsregeling, betalingsregelingen, bezwaarbehandeling, bezwaarschrift,

[1] Anthony, 2019

bezwaarschriften, bezwaarvoorziening, bijlage, bijlagen, bijtelling, boekenonderzoek, boekenonderzoeken, bpm, brexit, brief, brutocorrectie, bsn, btw, bulgarenfraude, burgerportal, burgerservicenummer, burgersvoor, bzm, ca, cafbestand, cafgerelateerd, cafll, cafzaak, cafzaken, campagne, cap, carrouselfraude, concerndirectie, continuÃ¯teitsrapportage, correctiebesluiten, correctieopbrengsten, cpb, dba, debiteuren, deurwaarderij, deurwaarders, dgb, dienstverlenersconvenant, dieseltoeslag, dieselvoertuigen, digitaal, dividendbelasting, domeinarchitecturen, dossier, douane, douanetaken, dwangbevelen, dwangbevelkosten, dwanginvordering, ecli, eenmanszaken, energie, ep, erfbelasting, erfbelastingsystemen, fec, fijnstofuitstoot, fiscale, forfaitair, fraude, fsv, ftedouane, ftefiod, fteiv, ftetoeslagen, ftetotaal, gaf, halfjaarrapportage, halfjaarsrapportage, halfjaarsrapportages, handhavingsbeleid, haventafel, hbb, heffen, heffingskortingen, huurtoeslag, ib, ih, ilt, inhoudingsplicht, inkomen, inkomensheffing, inkomensverklaring, inkomstenbelasting, innen, inningsverlies, inningsysteem, inspecteur, instroom, instroompercentage, integriteitinfrastructuur, investeringsagenda, invorderingen, invorderingsmaatregelen, invorderingsregelgeving, invorderingsregels, invorderingswet.

Institutions: abd, abdtopconsult, accenture, adr, agentschapsmodel, ambtsvoorganger, amsterdam, ap, ark, aruba, auditdienst, awir, awr, baliebezoekers, banken, basisregistratie, bd, bedrijven, belastingbetaler, belastingbetalers, belastingdienst, belastingdiensten, belastingdienstkantoren, belastingdienstmedewerkers, belastingen, belastingplichtige, belastingplichtigen, belastingschuldige, belastingtelefoon, beleidsdepartementen, belgiÃ«, bellers, berenschot, berichtenbox, bestuursrecht, bit, bonaire, broedkamer, bulgaren, bulgarije, burger, burgers, bzk, caf, caribisch, cbs, cda, college, committee, consumentenbond, curaÃ§ao, dg, dgbd, dgfz, dienstonderdeel, dienstonderdelen, directeurgeneraal, directoraat, douaniers, dr, dt, duitsland, ecd, eu, europa, eustatius, fd, financiÃ«n, fiod, fiscalisten, forum, fraudeteam, fz, gastouder, gastouderbureau, gastouderbureaus, gastouders, gemeentes, griffier, haag, handelsregister, houdstercoÃ¶peraties, huba, hypotheekhouder, hypotheekhouders, inhoudingsplichtige, inhoudingsplichtigen, inspecteurs, iv, jeugdorganisaties, justitie, kabinet, kamer, kamercommissie, kentekenhouder, kerndepartement, kinderopvang, kinderopvanginstelling, kinderopvanginstellingen, kinderopvangorganisatie, kinderopvangorganisaties, kinderopvangtoeslagaanvragers, koninkrijksrelaties, koophandel, landsadvocaat, lidstaten, lng, marechaussee, marokko, medewerkers, mevrouw, mijnbelastingdienst, mijnoverheid, minister, ministerie, ministeries, ministerraad, ministers, multinationals, nationaleombudsman, nederland, oeso, ombudsman, ondernemer, ondernemingsraad, ouders, overheidsondernemingen, parlement, particulieren, personeelsraadspersonen, politie, provinciale, provincies, psg, pvda, rdw, rechtspersonen, regering, rekenkamer, rijk, rijksbelastingen, rijksdienst, rijksoverheid, rotterdam, rvs, saba, secretaris, sg, softwareleveranciers, sp, sso, staatscourant, staatsecretaris, staatssecretaris, staatssecretarissen, suriname, szw, toeslagenstelsel, topstructuur, turkije, tweede, tweedekamer, ubo, uitvoeringsinstantie, uitvoeringsorganisatie, uitvoeringsorganisaties, uwv, vakbonden, vennootschappen, volksgezondheid, vraagouders, vvd, vws, waterstaat, werkgevers, wob, zzp.

Names: agnes, anne, appleby, azarkan, bas, bertholt, boonstra, brittanniÃ«, deloitte, eustasius, freriks, heeff, heerlen, hoekstra, huffelen, jaap, maassen, mulder, omtzigt, rutte, uijlenbroek, vries.

Formulas were created to remove jargon/name words from any newly added wordlist. This allows for analysis with and without the jargon/name words. In this way the following wordlists were created using AntConc[2]without sampling or culling)

a) List A1: full wordlist of 5000 most frequent words harvested across both parliamentary documents and the control set.
b) List A2: wordlist of 4629 most frequent words, harvested from list A1 above minus the list of jargon/name words
c) List B1: wordlist of 5000 most frequent words, harvested across parliamentary documents but not the control set.
d) List B2: wordlist of 4521 most frequent words, harvested from list B1 above minus the list of jargon/name words.

[2] Anthony, 2019

3 AUTHORSHIP ANALYSIS

The Stylo[3] package was used for authorship analysis. Stylo is freeware computer software built on top of R, a open-source statistical programming environment. Stylo allows for the processing of corpora consisting of many and/or large text large texts. It computes difference (distances) between text, represented as rows for frequencies of most frequent words. Then it plots graphs of these distances, so visualising the end result. For the computation of distances, classic delta (Burrows) was used.

The following graphs were used:

- Cluster analysis. This produces a dendogram which shows the similaries and dissimilarities between texts.
- Bootstrap cluster analysis: This produces a network diagram, based on the nearest relationships between texts. A network is produced for every MWF (most frequent word) setting; then all networks are combined into one consensus network
- Principal Component Analysis. This produces a scatter plot, showing the linear combination of variables (covariance) that best accounts for the variation in data.

[3] Eder, Rybicki, & Kestemont, 2016

3.1 INITIAL ANALYSIS

Using Stylo, a Cluster Analysis with classic Delta distance was run on corpus/wordlist B1 and B2, i.e. the primary set of document with and without jargon.

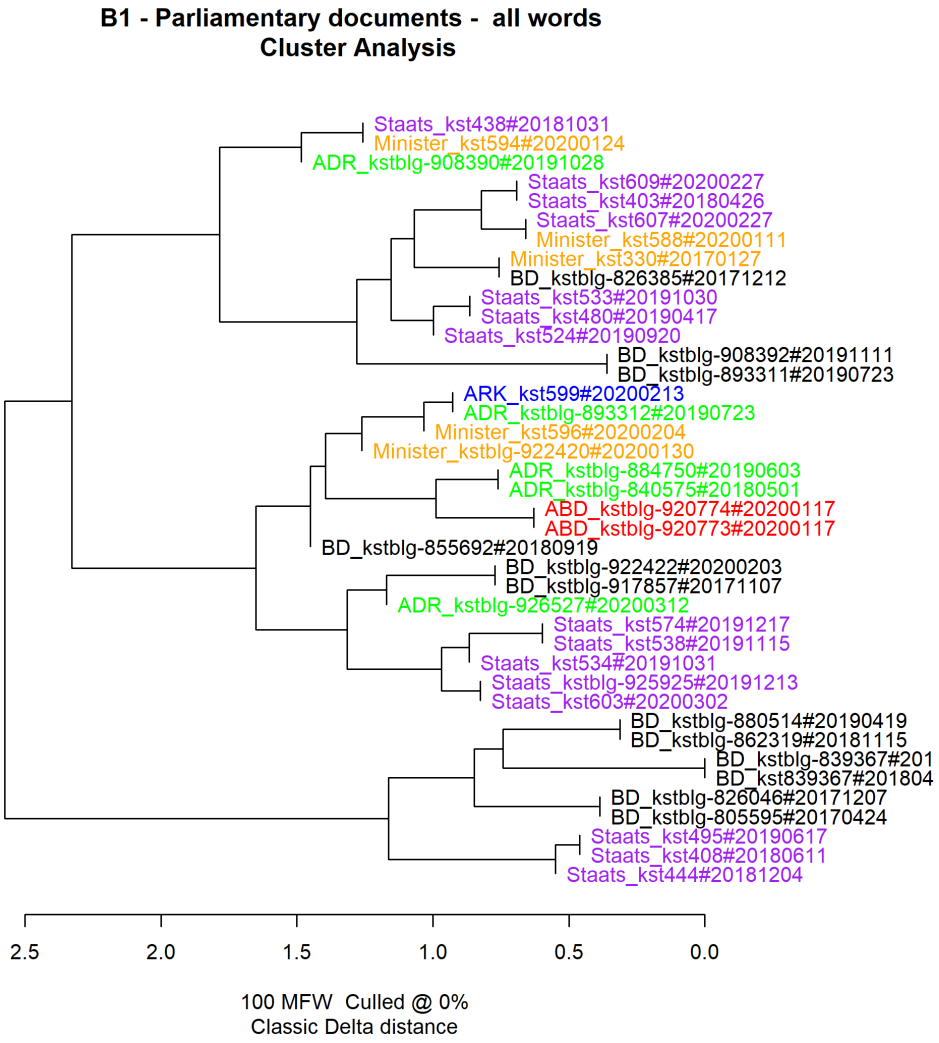


Figure 3 Stylo Cluster Analysis B1, 100 MWF, no sampling

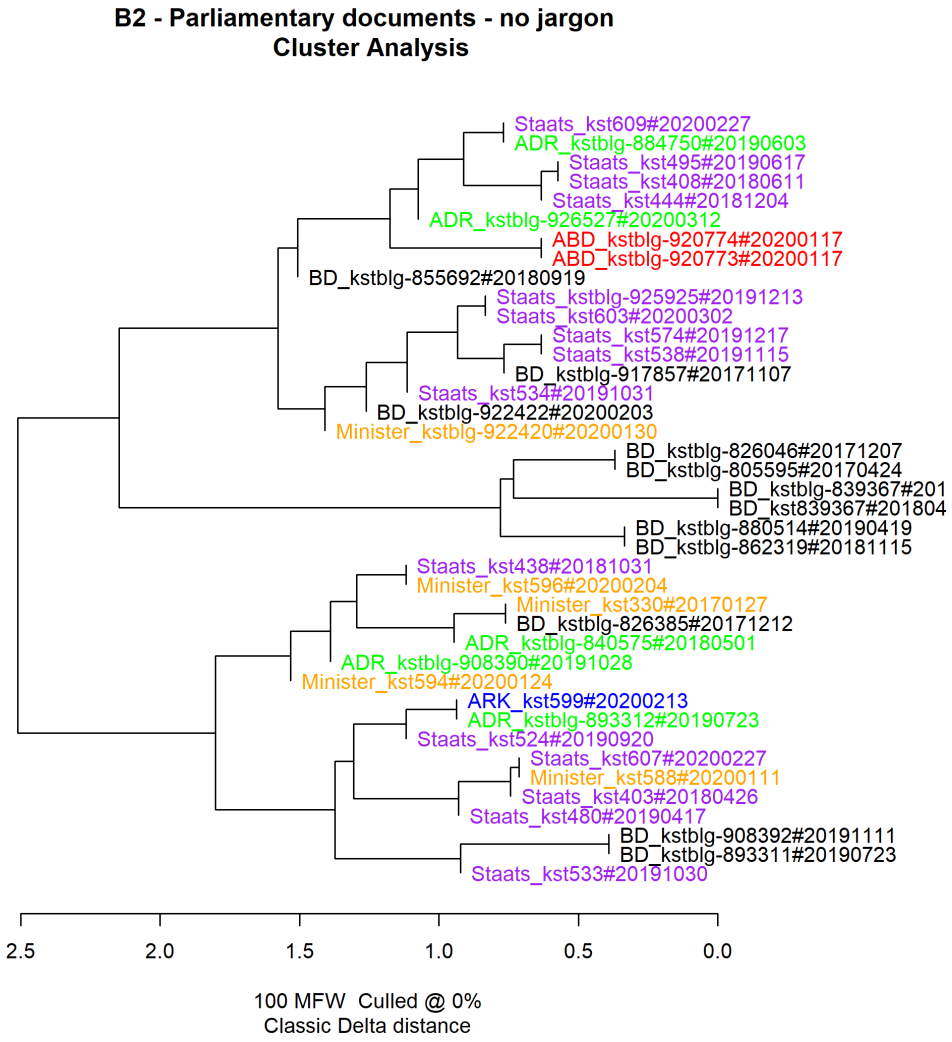


Figure 4 Stylo Cluster Analysis B2, 100 MWF, no sampling

The analysis was also run with and without deletion of pronouns. This made no difference, probably due to the peculiarities of modern official documents, these tend to be written in an unpersonal style.

Overall, there appear to be three groups:

a) Mainly consisting of documents by the Staatssecretaris (State Secretary) and the Auditdienst Rijk (Central Government Audit Service).
b) (Mainly) consisting of documents written by the Belastingdienst (Dutch Tax Office)
c) A mixed group, containing members of the previous groups plus the Minister, the Algemene Rekenkamer (General Audit Chamber) and Algemene BestuursDienst consultants (Civil Service)

On the B2 analysis (without jargon), the middle group (b) consists only of Belastingdienst documents. This effect is also produced when the mean word frequency is increased from 100 to 1500 in the B1 corpus/wordlist, as shown below. The shortest files produced the biggest outliers.

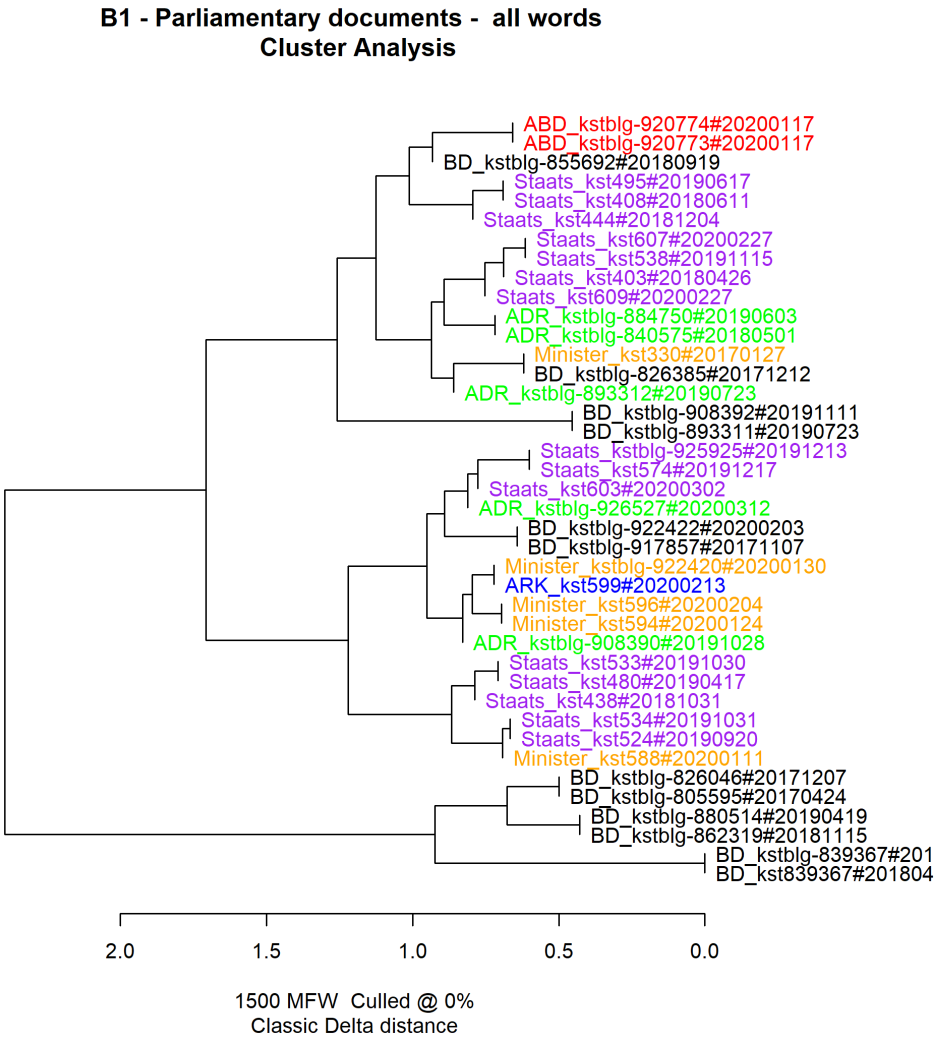


Figure 5 Stylo Cluster Analysis B1, 1500 MWF, no sampling

This result is supported by running the Bootstrap Consensus Tree which shows the middle group (b) Belastingdienst documents from corpus/wordlist B1 to belong to the same group as shown in the B2 Cluster analysis and the B2 Bootstrap consensus tree, as shown in the two graphs below.

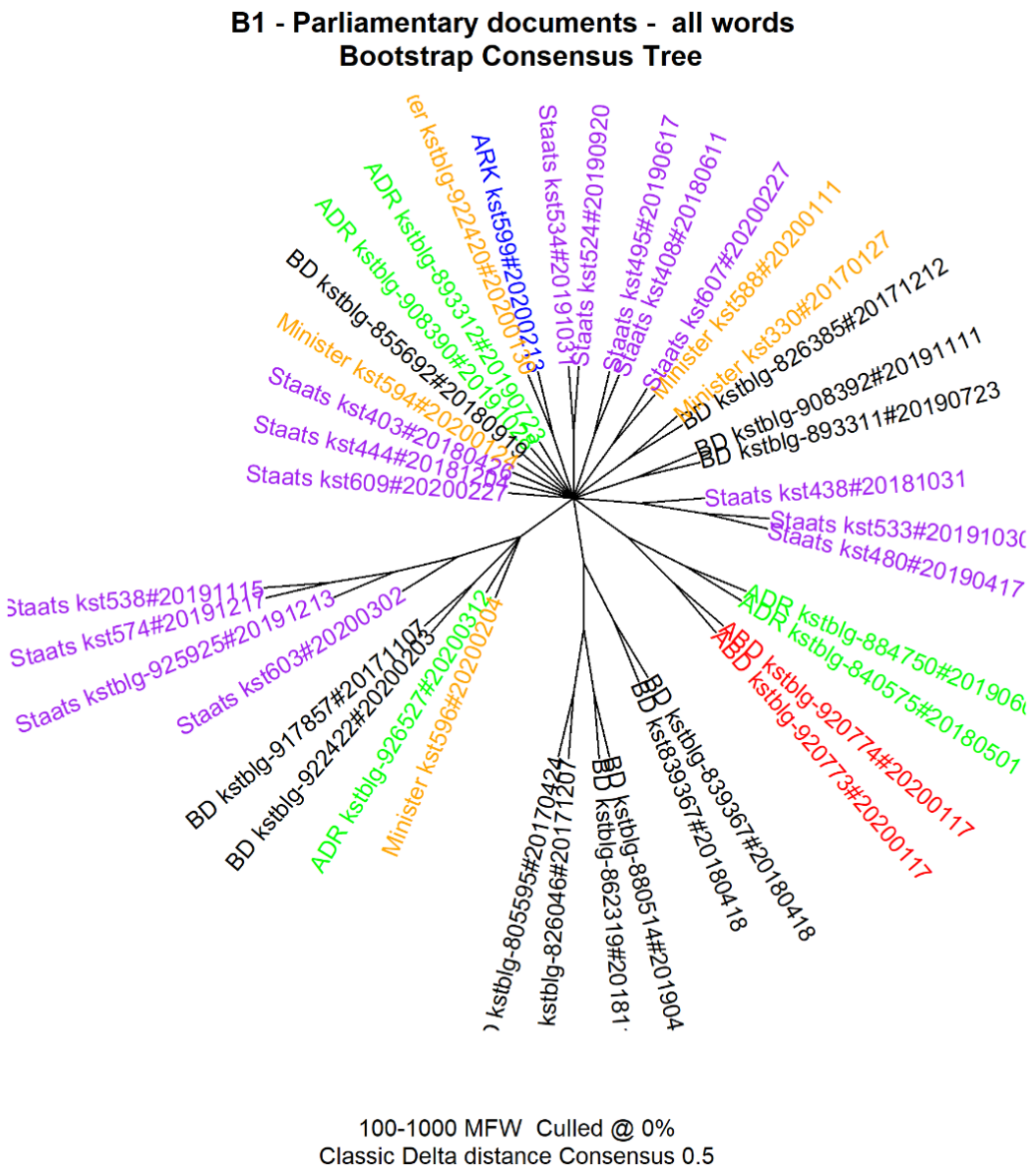


Figure 6 Bootstrap Consensus Tree B1, 100-1000 MWF, increment 100, no sampling

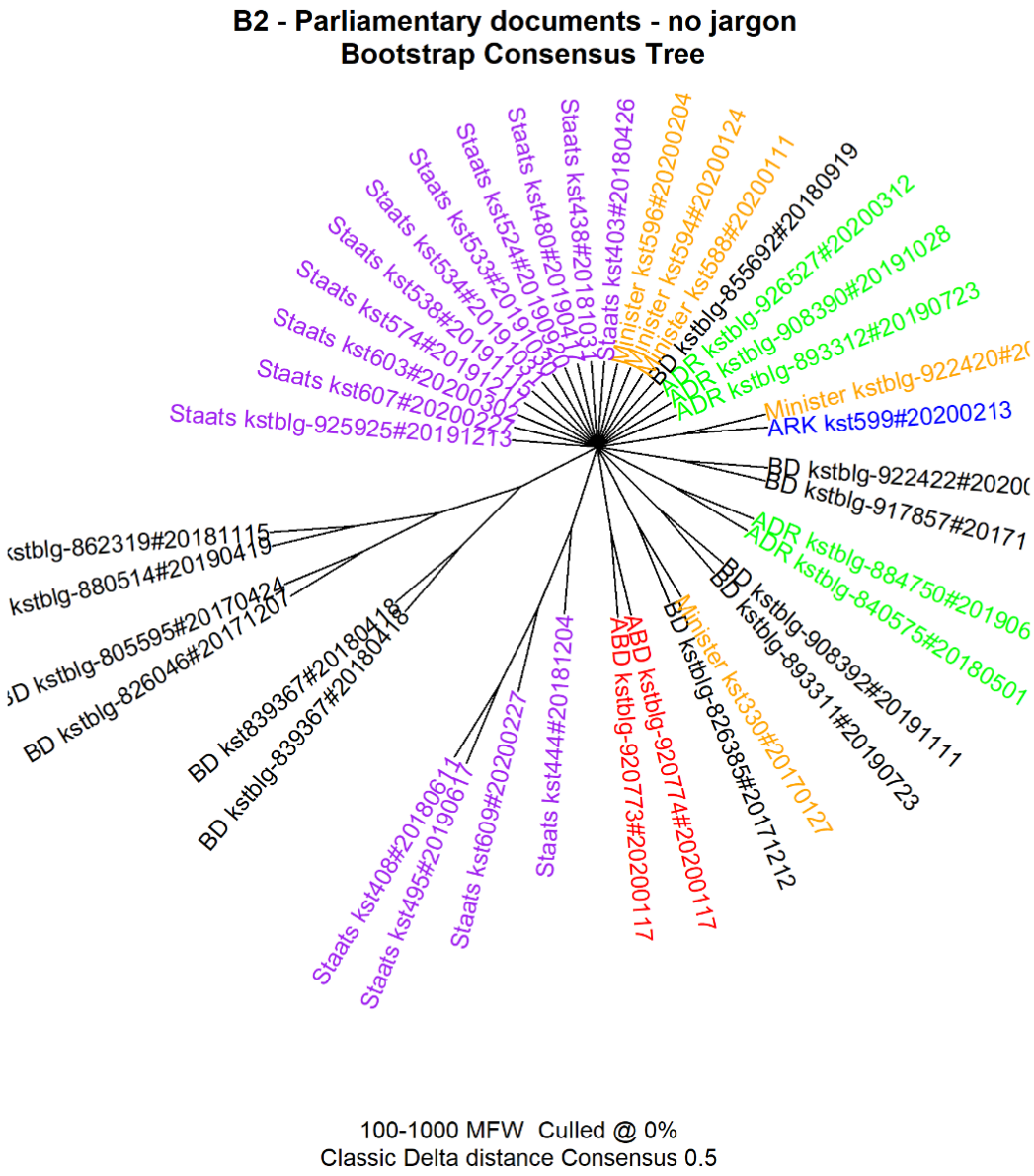


Figure 7 Bootstrap Consensus Tree B2, 100-1000 MWF, increment 100, no sampling

3.2 INTERPRETATION

These combined results indicate that:

- Documents Staats 408 and 395 were written by one author; together with documents Staats 609 and 444 they may have also been written by one author or a group of authors.
- Documents BD 862319 and 880514 were probably written by the same author; same for documents BD 805595 and 826046; and for documents BD 8393367 and its appendix; these six documents may also have been written by one author a group of authors.
- Documents BD 908392 and BD 893311 were written by the same person, but that person is not the author of the BD documents in the previous group.
- Documents BD 922422 and BD 917857 were by the same person, but that person is not the author of the BD documents in the previous group.

- Documents written on behalf of the Senior Civil Servant Consultants, ABD 920773 en ABD 920774, were written by one person.
- ADR documents 884750 and 84057 were written by the same person, but that person is not the author of the other documents written on behalf of the ADR.

So far, these results show a variety of authors who write on behalf of whatever party they are employed by. This is what was expected.

Most of the remaining documents appear to have been written by different authors:

- Most documents written on behalf of the State Secretary (except for cases Staats 408, 395, 609 and 444) were written by different authors who used a common jargon. This would explain the difference between the Bootstrap Consensus Tree B1 and B2 (purple font).
- Same for ADR documents 89331, 908390 and 926527 which appear to be have written by different authors. There is no influence from cutting out the jargon with these documents, i.e. the results on the Bootstrap Consensus Tree is the same on B1 and B2.
- Same for documents written on behalf of the Minister, documents 596, 594, and 588, again no influence from cutting out the jargon.

There are also some more mysterious results:

- Documents BD 826385 and Minister 330 appear to be written by the same author. This would mean that the Minister of Finance left the writing of this letter to the Dutch Tax Office, or alternatively, that one author wrote these documents – **and nothing else** in this set of documents. This seems unusual, as there is a huge hierarchical gap between the Dutch Tax office and the Minister of Finance, with several parties in between. Possibly this author, who must be quite knowledgeable to be writing on behalf of both parties, is someone from the Ministry of Finance, but if that is so, it is surprising that this person appears not to have written any other documents, for instance on behalf of the State Secretary (Staats). Possibly a larger corpus would shed light on this, but this is outside the scope of the current paper.
- Documents Minister 922420 and ARK 599 appear to be written by the same person. Both documents are about the so-called "Toeslagen affaire", parents persecuted for alleged fraud by the Dutch Tax office. The document from the General Audit Chamber is a critical report; the document by the minster (and the prime minister) a report on a between them and the affected parent. This overlap is remarkable, because the General Audit Chamber (ARK) audits the work that the Minister is responsible for; no contact would be expected, indeed would be frowned upon.

3.3 Result validation

So far, results from the set of parliamentary documents have been discussed. The question is whether the difference between these documents are significant. To check this, the A1 corpus was used, containing random documents from the internet and the set of parliamentary documents. Again, two versions were used for analysis, with and without jargon, as the highly specific jargon used in the parliamentary documents might create false negatives.

First, the results from corpus/wordlist A1. At 100 words, the Principle component analysis shows the parliamentary documents on the left side, the control set on the right side. The documents in the middle are two articles from Follow the Money, and the mystery document Minister 922420.

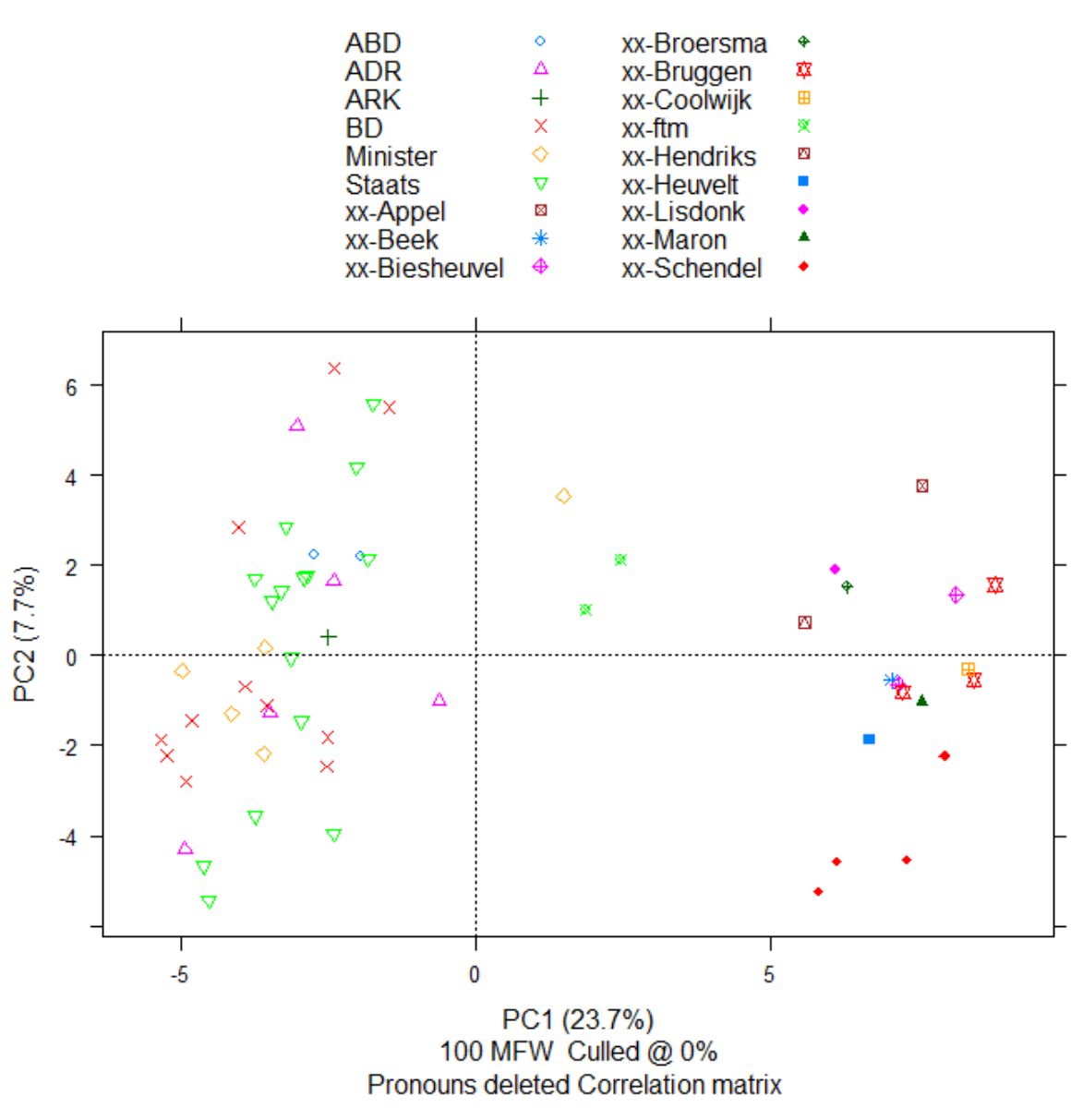


Figure 8 Principle Component Analysis A1, 100 MFW, no sampling, pronouns deleted

At 1500 MWF, the results are even more extreme, as shown below.

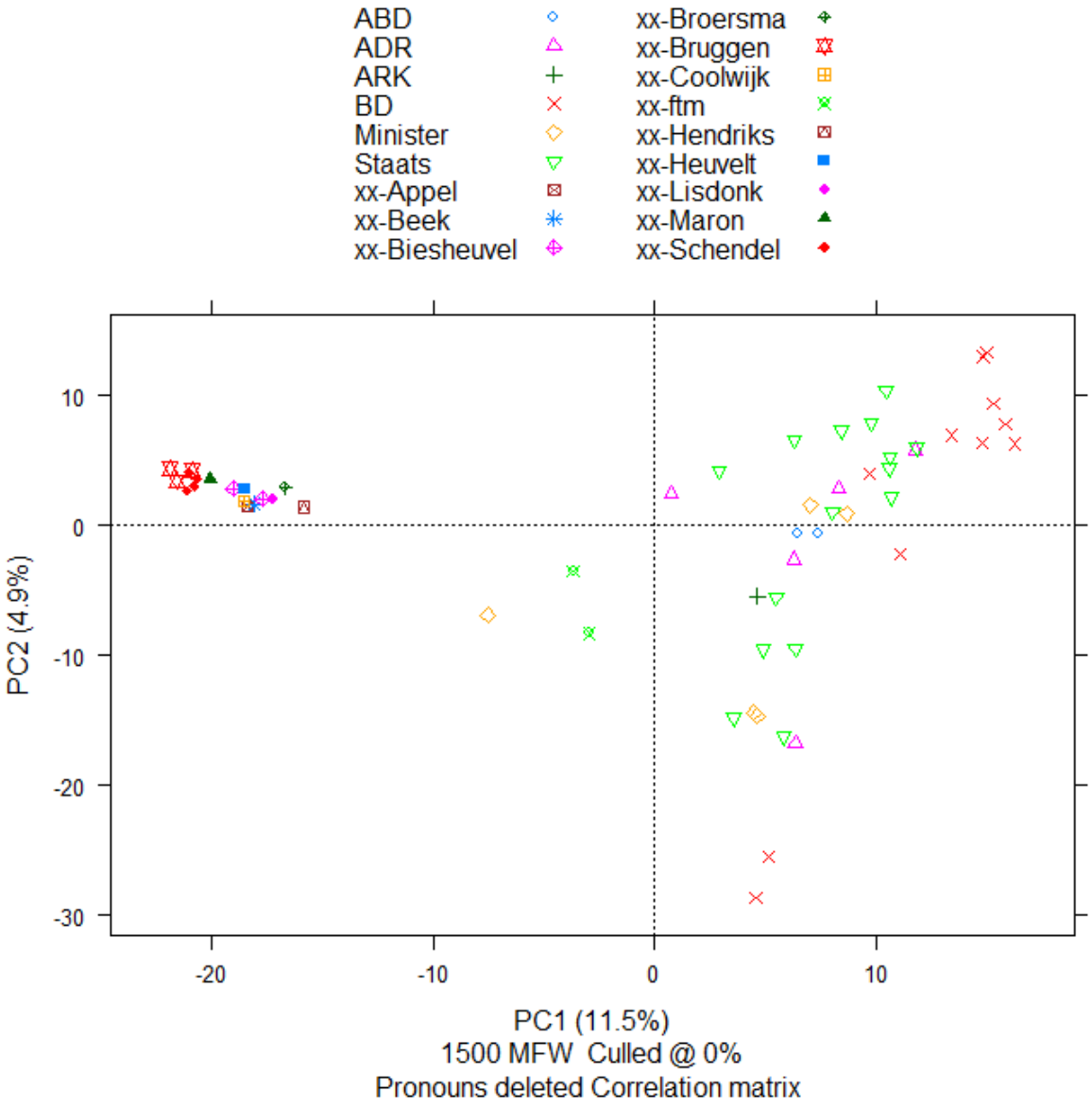


Figure 9 Principle Component Analysis A1, 1500 MFW, no sampling, pronouns deleted

Analysis of the A2 corpus/wordlist, without jargon, yields similar results at 100 MWF.

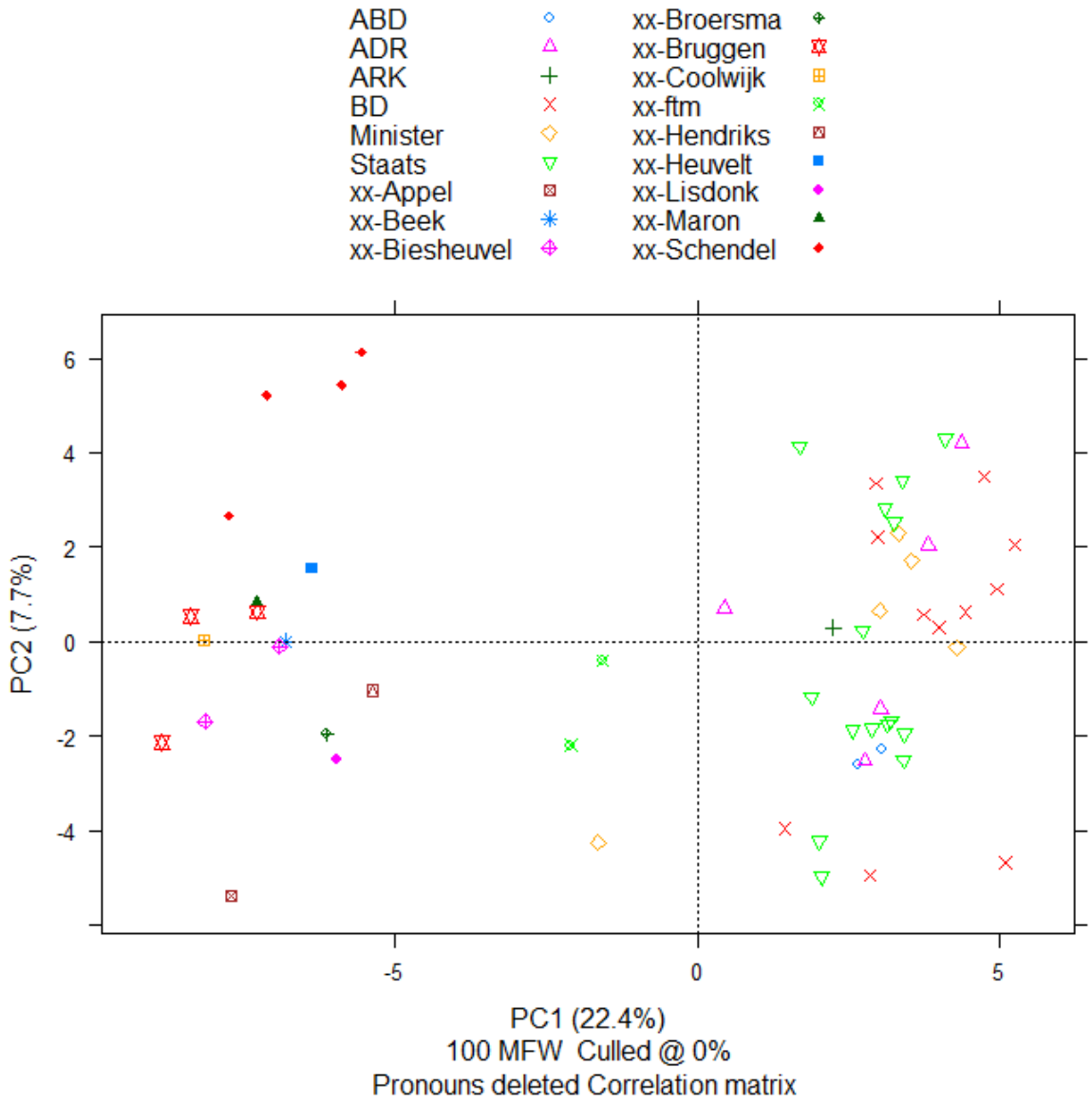


Figure 10 Principle Component Analysis A2, 100 MFW, no sampling, pronouns deleted

At 1500 MWF the result is more extreme, similar to the results with the A1 corpus/wordlist.

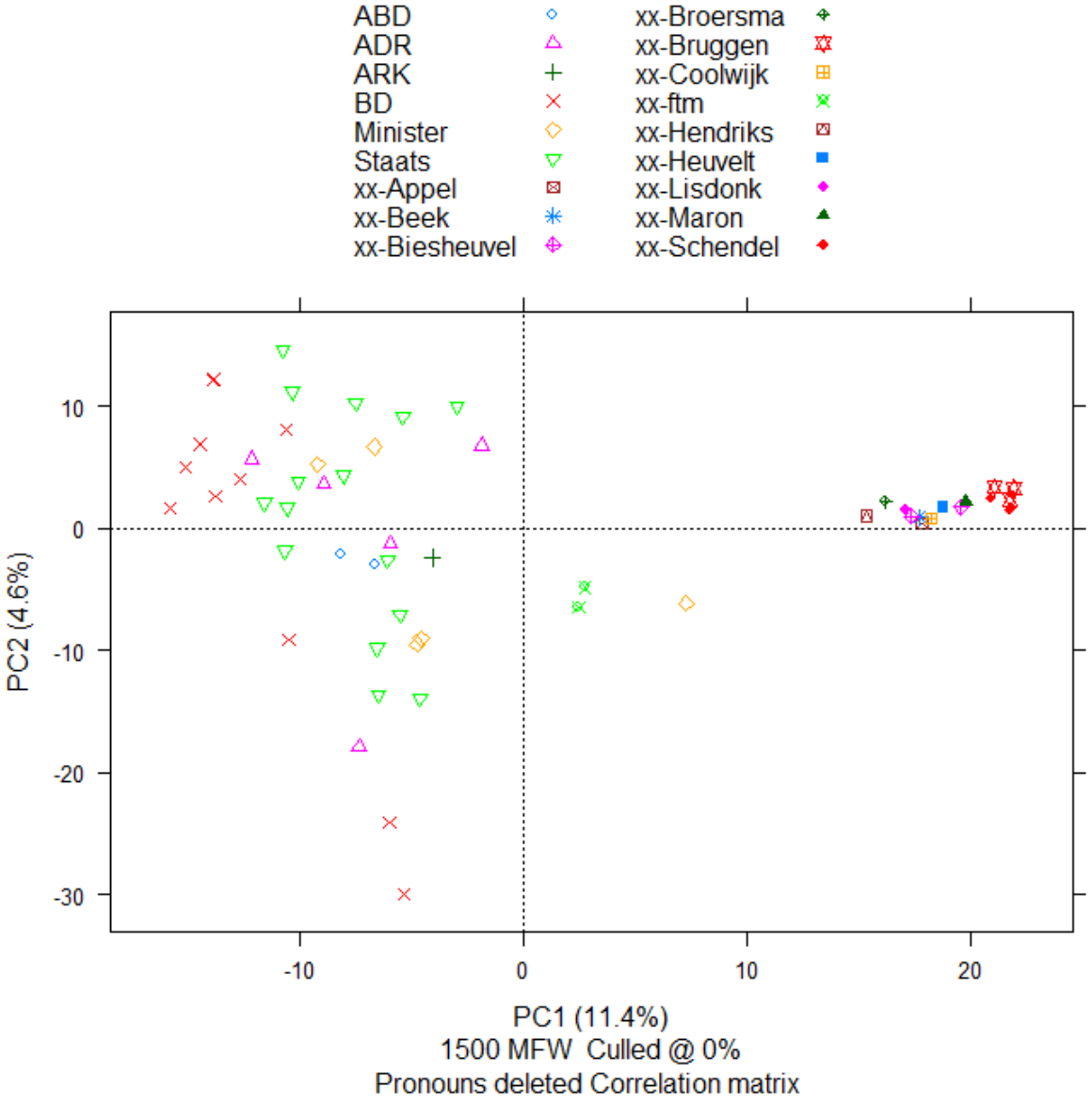


Figure 11 Principle Component Analysis A2, 1500 MFW, no sampling, pronouns deleted

From the analysis above, it is concluded that the set of parliamentary documents is quite different from the random control set, and therefore, that differences analysis between the parliamentary documents have significance.

3.4 Authorship conclusions

The assumptions arrived at in the previous paragraph were used to create a third corpus/wordlist B3. This corpus/wordlist is the same as B2, i.e. parliamentary documents without jargon, but with the names of the documents changed to reflect presumed authorship.

Running a cluster analysis with Stylo results in the following visualisation, with the presumed 27 author (groups) neatly separated out.

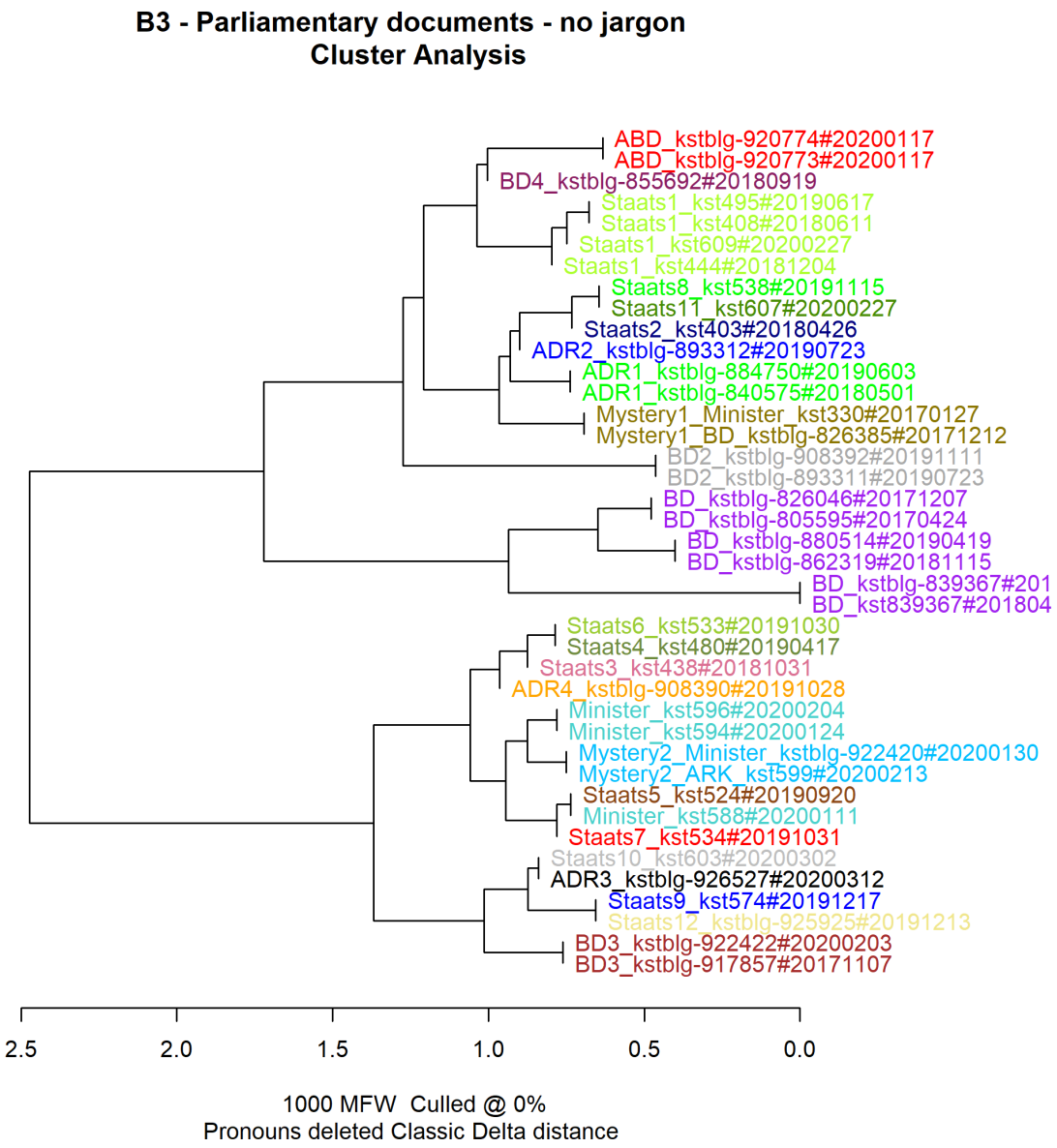


Figure 12 Stylo Cluster Analysis B4, 1000 MWF, no sampling, pronouns deleted.

Running the same analysis with pronouns not deleted or random sampling (500 words / 2) yields the same results.

The Bootstrap Consensus Tree now also shows the 27 presumed authors, as shown below:

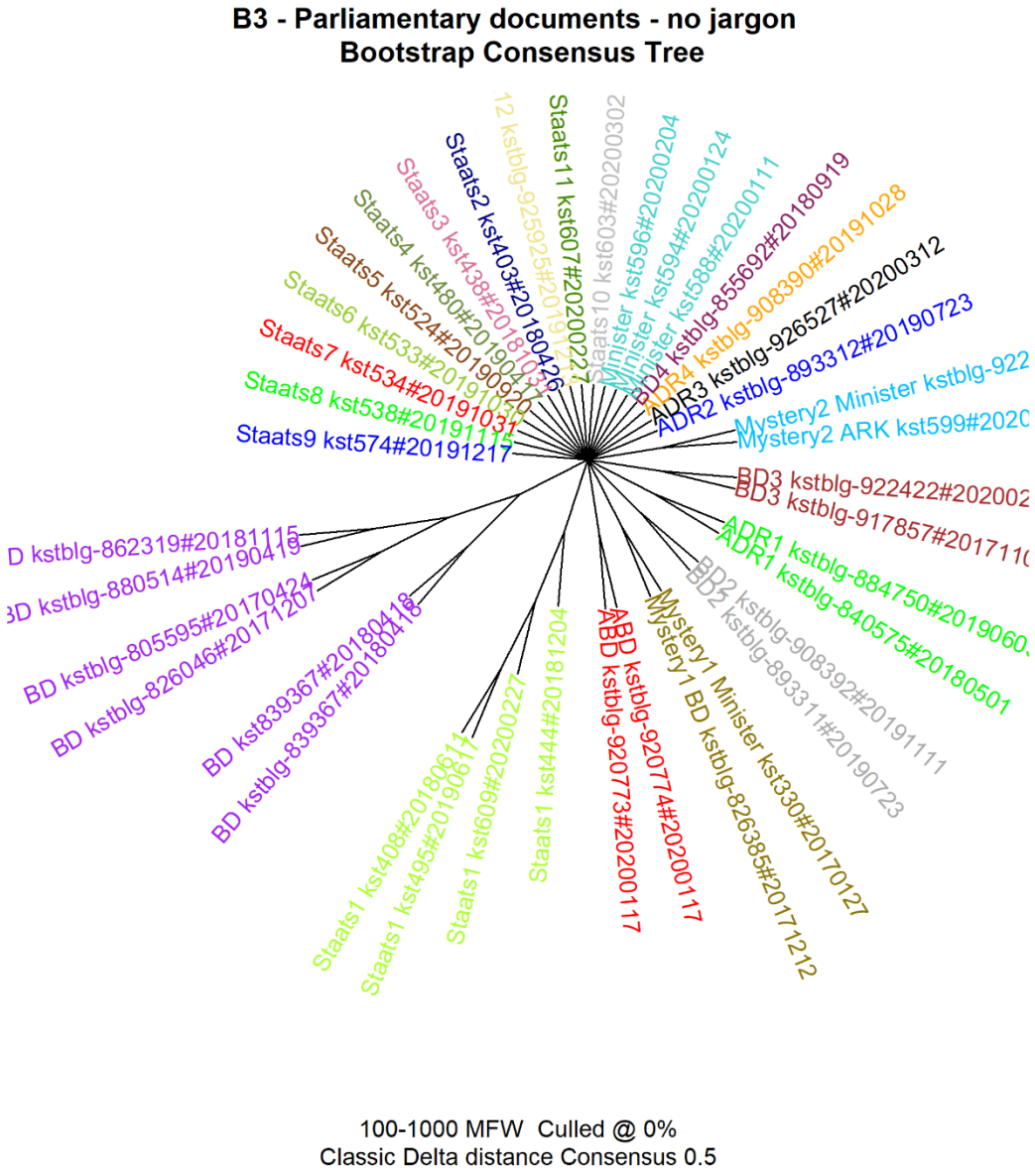


Figure 13 Bootstrap consensus tree, B3, increment 100, no sampling

It is not possible to analyse these data further using *Stylo*, 2016, because there is not one document for which authorship is known with any degree of certainty. The *Stylo*, 2016 functions ***oppose*** and ***rolling.classify*** and ***rolling.delta*** can therefore not be used.

When a much larger corpus is analysed, it may be possible to identify candidates-authors with which such analyses may be attempted.

4 DECEPTIVE LANGUAGE

The second part of the research question is about the extent to which deceptive language can be identified in parliamentary document. Newman, Pennebaker, Berry, & Richards, 2003 have established that deceptive language is characterised by:

- Less first- and third person pronouns
- More negative than positive emotion words
- More motion verbs
- More exclusion words

For this part of the analysis, LIWC2015[4] was used, the commercial version. LIWC2015 describes itself as the gold standard in computerised text analysis. It reads a given text and counts the percentage of words that reflects different emotions, thinking styles, social concerns and even parts of speech. LIWC has language specific dictionaries.

The parliamentary texts (corpus B3) were analysed using LIWC2015 as was the control set (corpus A), using the Dutch dictionary. The results are shown below.

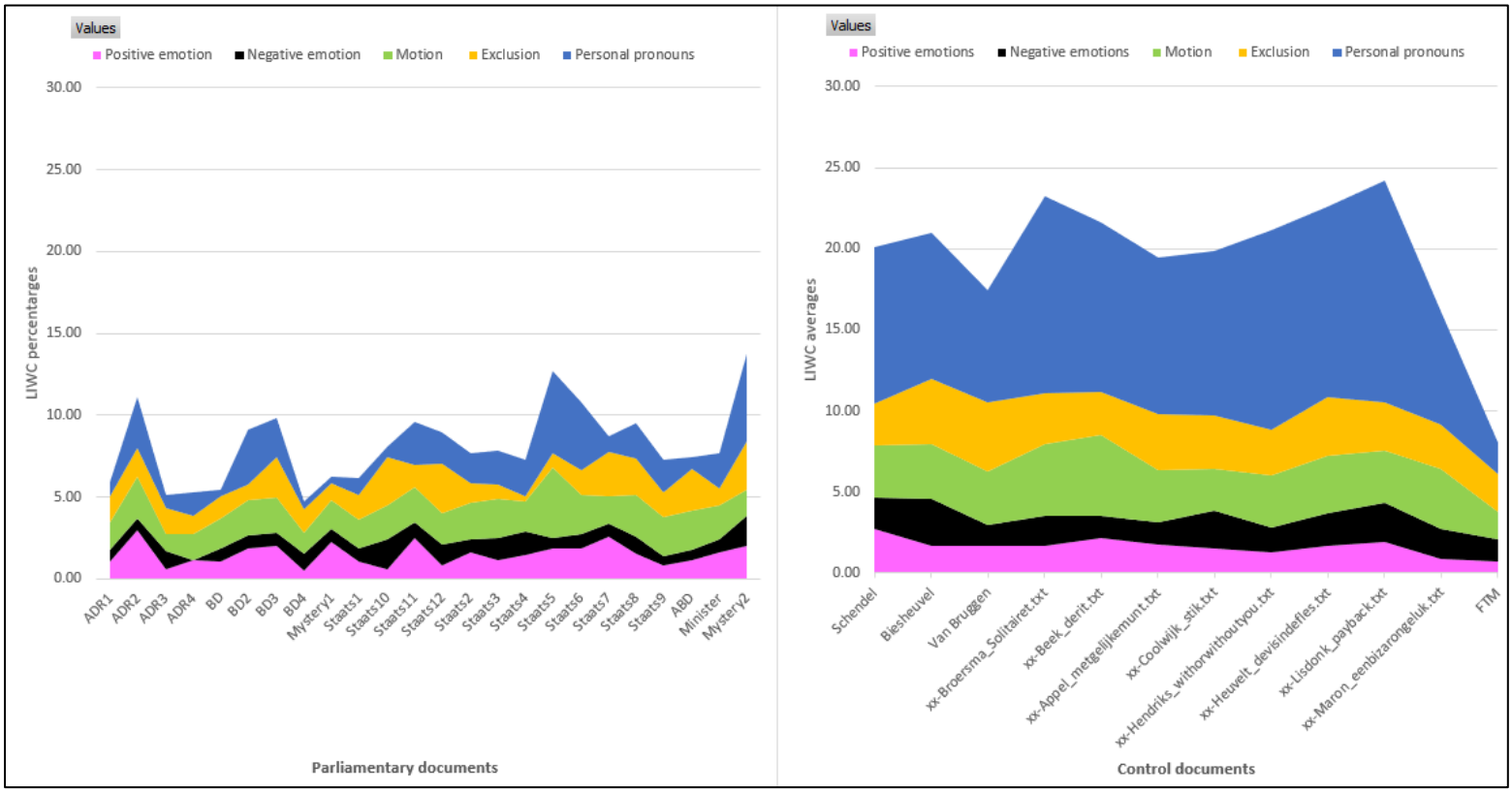


Figure 14 LWC comparison on deception markers

[4] Pennebaker, Boyd, Jordan, & Blackburn, 2015

Parliamentary documents contain very few personal pronouns as compared to the control set. However, this is unlikely to be an indication of deceptiveness, but more of distance (the author usually not the person on whose behalf the document is written) and of style. As for the other measures: *negative emotions* (as compared to positive emotions), *exclusion words* and *motion words*: parliamentary document contain less of these than the control set.

If these LICW indicators are good indicators of deceptiveness, then the conclusion must be that there are no traces of deceptiveness in parliamentary documents, as compared to the control set. Which is counter intuitive – we know from recent events that quite a few of the documents about the "Toeslagen affaire" contain falsehoods. However, the authors may not have known about this themselves, or not in much detail.

It is interesting to see that the FTM documents (newspaper) from the control set come out as similar to most parliamentary documents on these indicators. Possibly the LIWC indicators of deceptiveness are more about general style rather than about personal style.

Comparing the parliamentary documents and the control set on the other LWIC parameters, it appears that there are other differences in word use, as shown below:

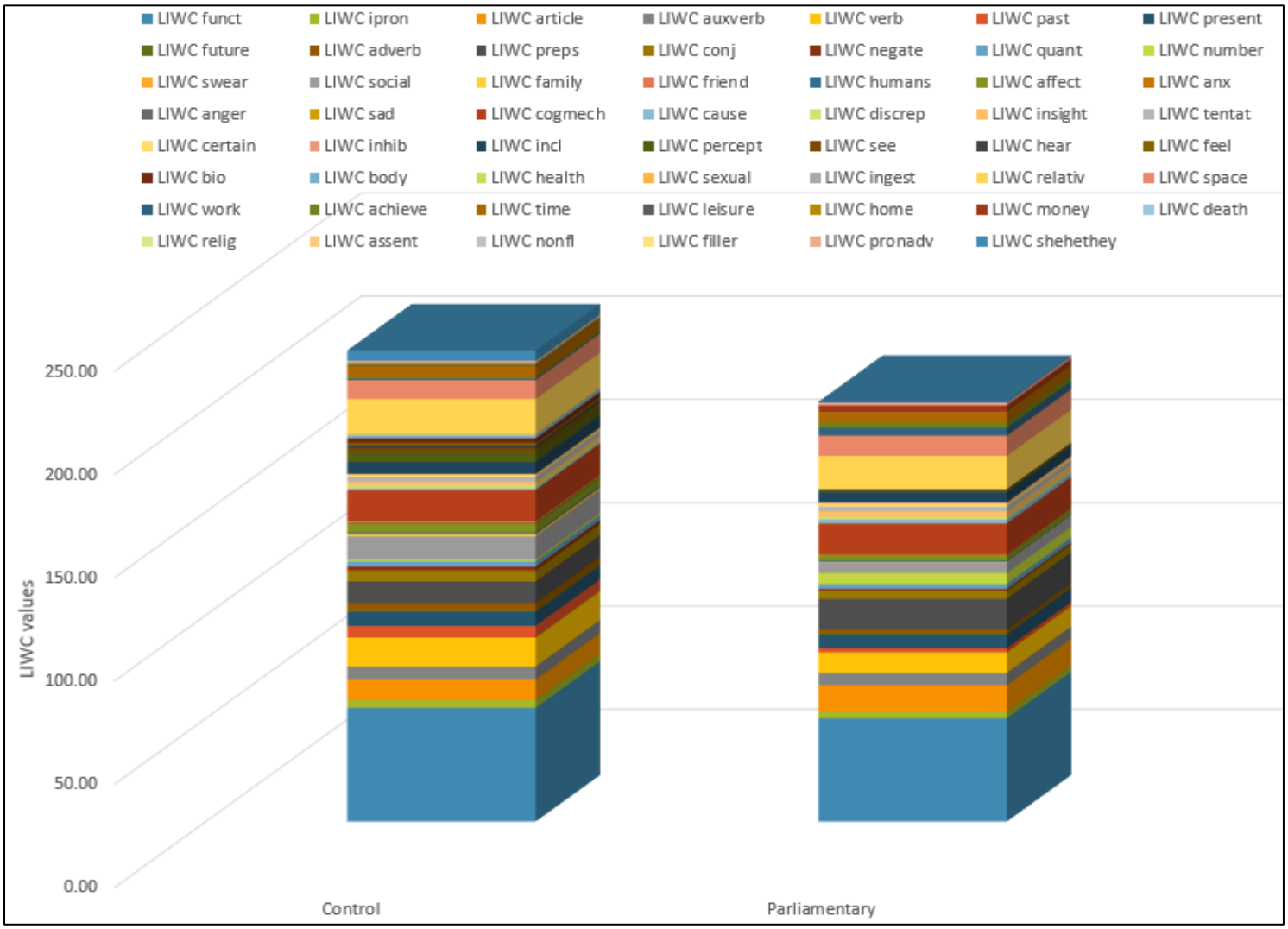


Figure 15 LIWC comparison on all parameters except deception markers and punctuation

The corpus of parliamentary documents contains:

- more words about *numbers, work, achievement,* and *money*
- fewer *social* words and *adverbs*, and words about *perception* and *bodily/biological* functions

as compared to the control group.

Again, this seems to be an indicator of general rather than a personal style. Presumably, there is also an influence from the subject matter.

5 Conclusion

This was a pilot study. The findings on authorship and deceptive language were presented in chapter 3.4 Authorship conclusions and chapter 4 Deceptive language. These findings would need to be confirmed in a larger study.

To make progress, several steps should be taken.

- A subsequent study should involve all documents from the dossier in a given period. A methodological decision should be made as to how to handle the difference in size – ranging from a few 100 words to a 100.000.
- This pilot study found style and jargon to be specific to parliamentary documents. The effects of these should be filtered out as early as possible. A methodological decision should be made as whether parliamentary documents belong to a genre or form a genre by themselves.
- Collect metadata on documents so that these can be used in further analysis. These should include: dossier, date, function, name, subject matter, type of document, reference to other documents. Only the first three parameters are standardised in the parliamentary dossier system, which is why only these were used in this pilot study. A classification system needs to be established for the other parameters. Once these parameters are known, it becomes possible to connect documents to each other, and understand the type of connection (standard report, answer to parliamentary questions, investigation etc). This may allow for better detection of defensiveness in texts, as a prelude to deception.
- It may be possible to obtain the names of the authors from the administration of the Ministry of Finance. Such a dossier does exist (Digidoc), but its accuracy is not known. It would be helpful in establishing likely authorship.
- A literature study should be done to find out if there are any other markers of deception apart from the ones investigated here. If not, then the problem should be approached backwards: work from known deceptive texts and try to establish markers.

6 Bibliography

Adobe Acrobat Pro DC Version Continuous release, version 2020.009.20065 [Windows]. (2020). Retrieved from https://acrobat.adobe.com/us/en/acrobat.html

Anthony, L. (2019). *AntConc (Version 3.5.8)[Computer Software]. Tokyo, Japan: Waseda University*. Retrieved from https://www.laurenceanthony.net/software

Count Anything Version 2.1 [Windows]. (2009). Retrieved from http://ginstrom.com/CountAnything/

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, *8*1, 107–121. https://doi.org/10/gghvwd

Microsoft Excel 365 Version Continuous [Windows]. (2020). Retrieved from www.office.com

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin*, *29*5, 665–675. https://doi.org/10/b8d7z6

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*.

Stylo. (2016). Retrieved from https://journal.r-project.org/archive/2016/RJ-2016-007/index.html