A diachronic corpus-assisted semantic domain analysis of US presidential debates

Nicholas Hayes[1] and Robert Poole[2]

Abstract

This corpus-assisted discourse study investigates diachronic change in a specialised corpus of seventeen American presidential debates from 2000 to 2020. The texts were tagged using the UCREL Semantic Annotation System (USAS) (Rayson, 2008) to facilitate the investigation of emergent and decreasing semantic trends over the period; the strength of trends was empirically evaluated through application of Kendall's Tau correlation coefficient. The analysis revealed that domains reflecting truth evaluations and matters of credibility increased alongside a more people-orientated discourse, as evidenced by increases in personal pronouns. Furthermore, instances invoking warfare and defence decreased, paralleled by a decrease in the representations of toughness. These results may reflect a shift in US political discourse generally, and American presidential discourse specifically, while also reflecting evolving contemporary social and political interests over the twenty-year span of the corpus. This study concludes with interpretations of these discursive shifts in the context of the current era of so-called 'fake news', intense partisanship, and social and political divisiveness. Findings indicate that the current US political climate cannot simply be attributed to an anomalous Trump administration but, rather, the discursive features contributing to and reflecting the current political environment have been present and increasing since at least the year 2000 election cycle.

Keywords: corpus-assisted discourse analysis, diachronic analysis, political discourse, presidential discourse.

[1] Clarendon Institute, Walton Street, Oxford, OX1 2HG, United Kingdom.
[2] Department of English, University of Alabama, 103 English Building, Box 870244, Tuscaloosa, AL 35487, USA.
Correspondence to: Nicholas Hayes, *e-mail*: nicholas.hayes@ling-phil.ox.ac.uk

Corpora 2022 Vol. 17 (3): 449–469
DOI: 10.3366/cor.2022.0266
© Edinburgh University Press
www.euppublishing.com/cor

1. Introduction

Corpus linguistics has long pursued investigations of diachronic language change. Traditionally, these investigations have focussed on the emergence and/or evolution of particular linguistic features across various periods of the English language. For example, such corpus-assisted diachronic research has investigated changing patterns of verb complementation in Late Modern English (Mair, 2002), the emergence of preposition patterns of pied piping and stranding in Middle English (Johansson, 2002), the changing use of modal verbs in the twentieth century in the US (Millar, 2009), and movement in English third-person singular present-tense suffixes from Late Middle to Early Modern English (Gries and Hilpert, 2010). These studies each explored large principled collections of historical language (i.e., corpora) in order to investigate grammatical changes across these historical periods. Such endeavours have been facilitated by the many historical corpora designed specifically for such investigations: the Helsinki Diachronic Corpus of English Texts (Rissanen *et al.*, 1991), the Archer Corpus of Historical English Registers (Biber and Finegan, 1990–) and the Corpus of Historical American English (Davies, 2010–), to name a few. Increasingly, however, diachronic corpus-assisted discourse studies (D-CADS) have explored language use and change more narrowly in specific discourses – for example, science communication (Poole *et al.*, 2019), academic writing (Hyland and Jiang, 2016), immigration discourse (Fitzsimmons-Doolan, 2019), treatment of anti-Semitism (Partington, 2012), changing representations of bisexuality (Wilkinson, 2019), evolving attitudes regarding morality (Marchi, 2010) and changing representations of species (Frayne, 2019). Perhaps claiming a social turn in D-CADS is not necessary, but it does appear evident that research in this space is more frequently exploring discourses of contemporary social import.

Similar D-CADS research, as the following review will demonstrate, have productively been pursued in political discourse analysis as well. That said, there are unique corpus analytic affordances which can yield insights into political discourse generally, and presidential debates specifically, that have yet to be implemented in this area. A diachronic semantic tag analysis is one such technique that has the capability to answer a range of research questions in political discourse analysis. The question this study poses is whether US political discourse of the Trump era was in fact anomalous and did indeed diverge from the conventions and norms of US political rhetoric. The linguistic variation of the Trump era is well-documented in existing research (Gonawela *et al.*, 2018; Clark and Grieve, 2019; and Ott and Dickinson, 2019) and has even been addressed within mainstream media (Golshan, 2016; Sedensky, 2017; and Wayne, 2017). This characterisation of former President Trump's communication as divergent seems somewhat common, and perhaps politically convenient, as President Biden assumes the role of president – the narrative that his administration returns the nation to a sense of normality seems pervasive. However, this study asks whether recent US presidential discourse actually represented such a stark transgression of discursive norms and conventions or whether the rhetoric of recent years reflects, reproduces and elaborates the discourse of contemporary US politics. To answer this question, this study performs a diachronic corpus-assisted analysis of semantic themes present in presidential debates from 2000 to 2020. The following section briefly reviews previous diachronic research of political discourse before introducing the affordances of the semantic domain analytic procedure and the quantitative analysis made possible through Kendall's Tau correlation coefficient.

2. Diachronic corpus-assisted study of political discourse

Researchers in political discourse analysis have increasingly applied corpus linguistic techniques to investigations that are both synchronic and diachronic in nature. This review focusses on studies interrogating diachronic change in political discourse. In one of the more comprehensive corpus-assisted studies in this field, a frequency analysis at the word level of inaugural and annual addresses from 1789 to 2000 revealed broad changes in language use (Lim, 2002). Namely, presidents have become decreasingly intellectual in their rhetoric, marked by a decrease in word references to cognitive, evaluative and stative processes (e.g., *effect*, *premise*, *consequence*, *authority*, *analysis*, *enact* and *administration*). This de-intellectualisation was joined by an increase in abstraction, with increasing word references to religious, poetic and idealistic constructs (e.g., *beauty*, *dream*, *faith* and *freedom*). Such abstract appeals, when coupled with increasingly people-orientated vocabulary (e.g., *our*, *family*, *children* and *friendship*), resulted in a more informal, conversational style, quantitatively supported by an increase in anecdotal features, descriptive verbs (e.g., *huddle*, *tell*, *call* and *gasp*), and comparable levels of first- and second-person singular pronouns. More specifically, usage of the inclusive self, *our*, has increased exponentially since Wilson's 1913 to 1921 presidency; the keywords *democracy* and *people* show increased usage from 1901 to 2000; appeals to the less fortunate have increased since 1933, marked by an increase of terms such as *poverty* and *help* – all indications of rhetoric that is focussed on establishing a connection with the audience (Lim, 2002). Lim's findings are consistent with those resulting from related methods in corpus linguistics. For instance, Tyrkkö's frequency analysis of pronoun usage in political speeches from 1800 to 2010 indicated a marked shift in inclusive pronouns (i.e., *we*, *our* and *us*) from the early 1900s onwards (Tyrkkö, 2016). Such work is further corroborated by El-Falaky's (2015) approach of using Halliday's systemic functional linguistics (SFL) to dissect the rhetoric of American presidential debates. Armed with a corpus of debates from 1960 to 2008, El-Falaky

(2015: 10) cited an increased usage in vocatives to assert that candidates ‘seek to sustain friendliness’ by invoking direct addresses and capturing the attention of voters. Perhaps surprisingly, El-Falaky noted that imperative clauses were less commonly employed than interrogative clauses; however, he also suggested that limited usage of imperative clauses is indicative of avoiding an authoritarian impression. Rather than giving orders, the candidate sought to establish an ‘equal and mutual reliant relationship’ wherein a candidate’s language use was reflective of this connection (El-Falaky, 2015: 6).

As suggested by Hart (1984), the presidency is changing in response to innovations in communicative technologies (e.g., television, radio and social media). To contextualise twenty-first century presidential debates by tracking the evolution of presidential rhetoric, it logically follows that discourses from the era of modern technology must be considered. Consider the 1996 presidential debates between Bill Clinton and Bob Dole, wherein Clinton, widely regarded as a master communicator, provided a sample of presidential rhetoric at the cusp of the twenty-first century. Clinton emerged from these occasions as the dominant rhetorician – a quality supplemented by evidence of comparatively greater use of vocatives, inclusive personal pronouns and well-crafted turn-initiators (Halmari, 2008). Halmari suggested that Clinton’s rhetoric is not just consistent with previous studies, but also helped define him as a superior debater. This viewpoint is further supported by a critical discourse analysis (CDA) of the 2004 George W. Bush and John Kerry debates. Jacobsen, narrowing in on question reformulation, similarly asserted that a forthcoming, well-received answer was one that responded to the formulation of the question rather than its implied meaning (Jacobsen, 2016). For the 2008 presidential campaign, the usage of the first-person plural *we* in prospective candidates can once again be observed, boosted by a contextual analysis from Proctor and Su (2011). Proctor and Su suggested that contextual surroundings of *we* usage provide valuable information about a candidate’s experiences and identity. Sarah Palin’s first-person plural pronouns were associated with common, middle class Americans, and Hillary Clinton’s first-person plural pronouns were associated with US government – an intuitive result considering her position as First Lady (Proctor *et al.*, 2011). Transitioning into the last decade, Wang and Liu’s (2018) language complexity approach to studying presidential rhetoric demonstrated how debate rhetoric is typically less complex than campaign speeches, citing Clinton and Trump’s 2016 and Obama’s 2012 presidential campaigns as evidence (Wang and Liu, 2018). Although the language of the American presidency is characterised by change, as each new office-holder inherently introduces a new linguistic style, the aforementioned works indicate a commonality among presidents: people-orientated discourse, defined by personal pronoun use and references to people and groups.

Corpus	No. of texts	No. of words
Presidential debates	17	231,603

Table 1: Presidential debates corpus.

While many previous studies have integrated corpus linguistic techniques to study political discourse (Lim, 2002; Baker and McEnery, 2005; Baker, 2012; El-Falaky, 2015; Romero *et al*., 2015; Tyrkkö, 2016; Aluthman, 2018; McDonnell, 2020; and Buckingham and Alali, 2019), this work has largely been pursued through the application and analysis of part-of-speech tags. Such studies of political discourse provide insight into the evolution of particular aspects of language; however, semantic annotation analysis can also provide alternative affordances for discourse analysis. Critical discourse analysis of a particular debate, such as Jacobsen's (2016) study of question reformulation in the 2004 debates, is successful in incorporating semantic analysis as a method of investigation but possesses shortcomings in scalability when a corpus is the subject of inquiry. This study attempts to extend aforementioned analytical techniques with semantic annotation analysis to address the following research question:

> What semantic domains and themes emerge and decline in presidential debates during the twenty-year period from 2000–2020?

3. Methodology

3.1 The Presidential Debates Corpus

The Presidential Debates (PD) corpus contains the three official, moderated debates from each presidential election from 2000 to 2020. Debate transcripts were accessed through UC Santa Barbara's *The American Presidency Project*, a non-partisan source of American presidential documents (University of California Santa Barbara, 2020). The debates were organised by year and by the two dominant political parties: Democratic and Republican. As third-party candidates infrequently participated in the debates, data for these candidates were not included, and utterances produced by moderators were removed from the texts as well. In addition, party-specific debates held during the respective primaries were not included. The architecture of the corpus enabled analysis across multiple parameters (e.g., diachronic comparative semantic tag analysis of Republican and Democratic candidates as well as diachronic semantic tag analysis of both parties). Details of the full PD Corpus as well as each election year sub-corpus are included in Tables 1 and 2.

Sub-Corpus	No. of texts	No. of words
2000 Election, Bush–Gore	3	39,844
2004 Election, Bush–Kerry	3	39,948
2008 Election, McCain–Obama	3	40,547
2012 Election, Romney–Obama	3	43,498
2016 Election, Trump–Clinton	3	40,024
2020 Election, Trump–Biden	*2	27,742

Table 2: Presidential debates sub-corpora. (*Only two debates occurred in 2020 due to the cancellation of the second scheduled debate.)

3.2 Semantic Analysis

Texts were annotated using the UCREL Semantic Analysis System (USAS) embedded within Wmatrix. The USAS tag system consists of twenty-one major semantic domains and 232 sub-domains (Wilson and Rayson, 1993). Examples of major semantic domains include, but are not limited to, education; money and commerce; government and the public domain; general and abstract terms; and even psychological actions, states, and processes. Examples of sub-domains within these categories include education in general; debts; warfare, defense, and the army; evaluation true/false; and conceptual objects, respectively. The annotation process assigns a unique semantic tag to all lexical items in the corpus which can subsequently be operationalised for various queries. The USAS system is sensitive to key relations between words (e.g., phrasal verbs, adverbial modifiers and negation), can identify idioms from a list of 16,411 items, and leverages syntactic tags to aid in word–sense disambiguation. The product of prior attempts to refine semantic analysis software (ACASD, ACAMRIT and REVERE), USAS unifies content analysis with automated approaches to text analysis (Rayson, 2002).

Informed by the diachronic analysis approaches of Hilpert and Gries (2009), Baron *et al.* (2009) and Poole *et al.* (2019), Kendall's Tau correlation coefficient was chosen as the measure to determine a time correlation in the normalised frequencies of sub-domains. Kendall's Tau is advantageous compared with more traditional measures of correlation, such as Pearson's r or Spearman's rho, because of its suitability for small sample sizes, ability to measure non-linear monotonic relationships and insensitivity to extreme values (Fredricks and Nelsen, 2007). Kendall's Tau is particularly well-suited to identifying gradual change, as it is sensitive to the direction of the change between election years (e.g., positive, negative or zero) rather than the magnitude of change. The numerical figure generated by Kendall's Tau ranges from –1 to +1; thus, interpretation is rather straightforward (Gries, 2010). For interpretation, scores approaching +1 represent positive correlations. In other words, a value closer to +1 reflects an increase in the use of a semantic category/tag over the time period under investigation. In contrast, scores trending toward –1 represent negative correlations – the use of the semantic category/tag decreased. The analysis focussed upon those semantic tags which exhibited a positive correlation statistic greater than or equal to 0.6, as well as semantic tags which displayed a negative correlation statistic less than or equal to –0.6. These cut-off points correspond to a statistical significance of $p < 0.1$. A cut-off point less than 0.1 was found to be too restrictive, whereas a cut-off point greater than 0.1 would not yield strong, convincing statistical significance. A cut-off of $p < 0.1$ was thus chosen to balance the issue of overlooking relevant data while maintaining statistical rigour. Semantic tags meeting these thresholds were further evaluated qualitatively by analysis of instances of their use in context.

4. Results

Correlating normalised semantic domain frequencies with the six election years denoted by the sub-corpora yields a total of forty-one domains with significant diachronic change. In other words, forty-one semantic domains met the aforementioned selection criterion for further analysis based on the strength of their correlation statistics. Fourteen of these domains demonstrate a positive correlation with time, meaning the normalised frequencies generally increase from 2000 to 2020. Conversely, twenty-seven sub-domains display a negative correlation with time, translating to a general decrease in frequency from 2000 to 2020. Table 3 lists those domains increasing over time whilst Table 4 lists those decreasing.

The first trend of significance is the perfect positive correlation of the ‘pronoun’ domain, represented by a maximum Kendall’s Tau value of +1. Pronoun usage has increased in each successive election year, with the largest change occurring as a nearly 15 percent increase from 2012 to 2016. Closer inspection reveals that pronoun usage is dominated by the first-person *I* and *we* pronouns, accounting for 16 percent and 14 percent of all pronouns, respectively. The increase in all pronouns is a result of both political parties; however, when calculating Kendall’s Tau values for the normalised frequencies when separated by party, the Republicans (Tau = +0.6, $p = 0.068$) demonstrate a much greater positive correlation than the Democrats (Tau = +0.333, $p = 0.2345$). Based on the associated p-values with these correlation statistics, the data suggest that the Republican Party dominates the increase in pronouns, whereas the weak positive correlation present in the Democratic party is not statistically significant. While this may indicate an increasingly people-orientated discourse, consistent with existing literature, it appears to be counter-balanced by a decreasing trend in the ‘groups and affiliation’ sub-domain, which yields a Tau value of –0.867 and a maximum 21 percent decrease from 2000 to 2004. The ‘groups and affiliation’ domain may initially be expected to be referring to groups of people, yet words tagged in this domain are rather referring to organisational

Tag	Domain	2000	2004	2008	2012	2016	2020	Kendall's Tau	*p*-value
Z8	Pronouns	14,562	15,095	15,271	15,302	17,584	18,618	1	0.0015
A13.1	Degree: Non-specific	25	65	37	69	115	130	0.867	0.0085
Q2.1	Speech etc: Communicative	645	1,064	952	1,081	1,164	1,409	0.867	0.0085
T1.1.3	Time: General: Future	1,062	1,136	1,216	1,287	1,219	1,373	0.867	0.0085
X3.4	Sensory: Sight	146	210	274	317	490	332	0.867	0.0085
A13.7	Degree: Minimizers	8	23	12	16	27	50	0.733	0.028
M1	Moving, coming and going	710	914	725	839	972	1,107	0.733	0.028
T2	Time: beginning and ending	429	463	506	635	515	616	0.733	0.028
W2	Light	3	3	7	5	12	18	0.733	0.028
A13.4	Degree: Approximations	80	60	72	80	112	87	0.6	0.068
A5.1	Evaluation: Good/Bad	760	706	814	802	1,509	966	0.6	0.068
A5.2	Evaluation: True/False	173	243	232	228	337	562	0.6	0.068
M5	Movement/transportation: air	8	25	22	23	32	25	0.6	0.068
O4.3	Color and color patterns	10	33	17	34	30	115	0.6	0.068

Table 3: Positively correlated semantic domain frequencies (normalised per 100,000 words).

Tag	Domain	2000	2004	2008	2012	2016	2020	Kendall’s Tau	*p*-value
S1.2.5	Toughness; strong/weak	198	193	148	168	122	50	–0.867	0.0085
S5	Groups and affiliation	615	488	427	453	387	328	–0.867	0.0085
X6	Deciding	161	188	99	97	95	47	–0.867	0.0085
Z5	Grammatical bin	28,351	28,487	28,046	28,045	26,524	24,937	–0.867	0.0085
A1.5.1	Using	156	125	86	62	105	47	–0.733	0.028
G3	Warfare, defense, and the army; Weapons	512	916	439	439	402	205	–0.733	0.028
O4.5	Texture	50	38	22	18	32	7	–0.733	0.028
S8	Helping/hindering	951	776	819	609	675	415	–0.733	0.028
X2	Mental actions and processes	13	10	7	9	2	4	–0.733	0.028
X2.1	Thought, belief	1,082	801	604	474	762	386	–0.733	0.028
X4.2	Mental object: means, method	319	285	239	278	245	155	–0.733	0.028
X5.1	Attention	68	38	59	51	17	11	–0.733	0.028
X7	Wanting; planning; choosing	1,272	976	1,134	1,120	867	764	–0.733	0.028
A1.3	Caution	35	28	15	16	17	14	–0.6	0.068
A11.1	Importance: Important	452	280	545	290	270	231	–0.6	0.068
A5.3	Evaluation: Accuracy	238	325	183	200	187	141	–0.6	0.068
A6.1	Comparing: Similar/different	718	566	629	526	585	519	–0.6	0.068
A6.3	Comparing: Variety	30	30	7	14	17	4	–0.6	0.068

Table 4 (*continued on following page*): Negatively correlated semantic domain frequencies (normalised per 100,000 words).

Tag	Domain	2000	2004	2008	2012	2016	2020	Kendall's Tau	*p*-value
C1	Arts and crafts	40	50	27	39	32	25	–0.6	0.068
N3.1	Measurement: General	50	25	10	7	2	11	–0.6	0.068
N3.7	Measurement: length and height	138	93	111	136	77	68	–0.6	0.068
P1	Education in general	781	323	291	506	155	260	–0.6	0.068
Q1.1	Communication in general	148	163	109	99	115	54	–0.6	0.068
S7.1	Power, organizing	1,144	668	592	715	367	552	–0.6	0.068
T1.1.2	Time: General: Present: simultaneous	532	596	614	492	397	360	–0.6	0.068
W1	The universe	153	270	153	228	117	83	–0.6	0.068
X2.5	Understand	146	168	192	101	80	61	–0.6	0.068

Table 4 (*continued from previous page*): Negatively correlated semantic domain frequencies (normalised per 100,000 words).

1. 10/3/2000, Bush	Actually what I've **said**, Jim. I've **said** that eight years ago they campaigned on prescription drugs for seniors. And four years ago they campaigned on getting prescription drugs for seniors. And now they're campaigning on getting prescription drugs for seniors. It seems like they can't get it done
2. 10/8/2004, Bush	… it's a fundamental misunderstanding to **say** that the war on terror is only Usama bin Laden. The war on terror is to make sure that these terrorist organizations do not end up with weapons of mass destruction. That's what the war on terror is about.
3. 09/30/2004, Kerry	You **talk** about mixed messages. We're **telling** other people, 'You can't have nuclear weapons,' but we're pursuing a new nuclear weapon that we might even contemplate using. Not this President. I'm going to shut that program down, and we're going to make it clear to the world we're serious about containing nuclear proliferation
4. 10/9/2016, Clinton	And, you know, Donald **says** he knows more about ISIS than the generals. No, he doesn't.
5. 10/3/2012, Romney	First of all, I don't have a $5 trillion tax cut. I don't have a tax cut of the scale that you're **talking** about. My view is that we ought to provide tax relief to people in the middle class.
6. 10/16/2012, Romney, Obama	*Romney*: Production on Government land of oil is down 14 percent and production of gas is down 9 percent. *Obama*: What you're **saying** is just not true. It's just not true.

Table 5: Excerpts from 'Q2.1 speech: communicative' (MM/DD/YYYY).

groups. In fact, the word *federal* is the top contributor to this domain at over 10 percent. Other examples include *allies*, *group*, *middle class* and *organization*. As such, this domain does not necessarily oppose the idea of increasingly people-orientated semantic representations.

Among the sub-domains similarly demonstrating nearly perfect positive correlations, the coupled 'speech: communicative' and 'time: general: future' sub-domains suggest increasing references to promises, pledges and commitments, particularly in the future. Words tagged in the 'speech: communicative' sub-domain are often allusions to another candidate or one's own words, followed by an evaluation of whether the verbal commitment was acted upon (see Table 5, Examples 1, 3, 5). In fact, the words of quotation (*said*, *say*, *talk*, *says*, *talking* and *saying*) make up 73 percent of all words tagged in this sub-domain. When looking at this sub-domain along with 'time: general: future', these sub-domains demonstrate a particular increase between 2016 and 2020, experiencing their largest increases of 13 percent and 21 percent, respectively. The idea of increasingly frequent promises, pledges and commitments is further substantiated by a positive trend in the 'time: beginning and ending' sub-domain (Tau = +0.733), often referencing the end of one politician's policy and the promise of a new era (see Table 6).

The most intriguing of results is the strong positive correlation found in the 'evaluation: true/false' sub-domain alongside an equally strong

1. 10/22/2020, Biden	Well, if you let me finish the statement, because it has to be replaced by renewable energy over time. Over time. And I'd **stop** giving to the oil industry-- I'd **stop** giving them federal subsidies.
2. 09/29/2020, Trump	**By the end** of the first term, I'll have approximately 300 Federal judges and Court of Appeals judges, 300, and hopefully three great Supreme Court judges, justices.
3. 10/19/2016, Trump	We are going to make America strong again, and we are going to make America great again, an d it has to **start** now. We can not take four more years of Barack Obama, and that's what you get when you get her.
4. 10/7/2008, Obama	This is not the **end** of the process; this is the **beginning** of the process. And that's why it's going to be so important for us to work with homeowners to make sure that they can stay in their homes. The secretary already has the power to do that in the rescue package, but it hasn't been exercised yet. And the next president has to make sure that the next Treasury secretary is thinking about how to strengthen you as a home buyer, you as a homeowner, and not simply think about bailing out banks on Wall Street.
5. 10/7/2008, Obama	You know, you may have seen your health care premiums go up. We've got to reform health care to help you and your budget. We are going to have to deal with energy because we can't **keep on** borrowing from the Chinese and sending money to Saudi Arabia. We are mortgaging our children's future. We've got to have a different energy plan. We've got to invest in college affordability.

Table 6: Excerpts from 'T2 time: beginning and ending' (MM/DD/YYYY).

negative correlation in the 'evaluation: accuracy' sub-domain. In an era of so-called 'fake news' and accusations of credibility, it may be supposed that 2020 yields a higher frequency in the 'evaluation: true/false' sub-domain. This is indeed the case, with 2020 possessing the highest normalised frequency following a 67 percent increase from 2016. Despite this, the data suggest that the elevated frequency in 2020 is the culmination of a positive trend since 2000 (Tau = +0.6). Although the highest frequency of the 'evaluation: true/false' domain occurs in 2020, this increase is the pinnacle of twenty years of evidence pointing towards an increase in truth evaluations. Words tagged in this domain are often attacking an opponent's credibility, correcting an opponent's statement, or reaffirming one's own position as the true, superior perspective (see Table 7, Examples 2, 4, 6, 8, 10). Initial analysis of the strong negative correlation (Tau = –0.6) found in 'evaluation: accuracy' seems to contradict these results; however, closer inspection affirms the idea of candidates questioning credibility. The most frequent occurrences within 'evaluation: accuracy' are indications of

1. 09/29/2020, Biden	First of all, that’s simply not **true** what he just said, of course
2. 10/17/2000, Bush	Actually, Mr. Vice President, it’s not **true**. I do support a national patient’s bill of rights. As a matter of **fact**, I brought Republicans and Democrats together to do just that in the State of Texas to get a patient’s bill of rights through
3. 09/30/2004, Kerry	I mean, this is the President who said there were weapons of mass destruction, said ‘mission accomplished, ‘said we could fight the war on the cheap, none of which were **true**.
4. 10/8/2004, Kerry	Ladies and gentlemen, that’s just not **true**, what he said. The Wall Street Journal said 96 percent of small businesses are not affected at all by my plan.
5. 10/13/2004, Bush	Well, first of all, it is just not **true** that I haven’t met with the Black Congressional Caucus.
6. 10/3/2000, Bush	You know, this man has no **credibility**on the issue. As a matter of **fact**, I read in the ‘New York Times’ where he said he co-sponsored the McCain-Feingold Campaign Fundraising Bill. But he wasn’t in the Senate with Senator Feingold
7. 09/30/2004, Kerry	Now, we can succeed, but I don’t believe this President can. I think we need a President who has the **credibility** to bring the allies back to the table and to do what’s necessary to make it so America isn’t doing this alone
8. 09/30/2004, Kerry	I will bring fresh **credibility**, a new start, and we will get the job done right.
9. 10/9/2016, Clinton	And he never apologized for the racist **lie** that President Obama was not born in the United States of America.
10. 10/9/2016, Trump	You know that, because Jonathan Gruber, the architect of Obamacare, was said he said it was a great **lie**, it was a big **lie**
11. 10/19/2016, Trump	One **lie**. She’s **lied** hundreds of times to the people, to Congress, and to the FBI. He’s going to probably go to jail. This is a four-star general. And she gets away with it, and she can run for the presidency of the United States?
12. 10/9/2016, Clinton	But I think it’s also important to point out where there are some **misleading** accusations from critics and others. After a year-long investigation, there is no evidence that anyone hacked the server I was using and there is no evidence that anyone can point to at all. Anyone who says otherwise has no basis that any classified material ended up in the wrong hands. I take classified materials very seriously and always have
13. 09/26/2016, Clinton	Donald thinks that climate change is a **hoax** perpetrated by the Chinese. I think it’s real.

Table 7: Excerpts from ‘A5.2 evaluation: true/false’ (MM/DD/YYYY).

1. 10/11/2000, Gore	I speculate that the reason why he didn’t answer your question directly as to whether my numbers were **right**, the facts were **right** about Texas ranking dead last in families with health insurance and 49th out of 50 for both children and women, is because those facts are **correct**
2. 10/8/2004, Kerry	Iran/North Korea I don’t think you can just rely on U.N. sanctions, Randee, but you’re absolutely **correct**. It is a threat
3. 10/13/2004, Kerry	*Moderator*: You’re also talking about the Government picking up a big part of the catastrophic bills that people get at the hospital. And you have said that you can pay for this by rolling back the President's tax cut on the upper two percent. *Kerry*: That’s **correct**.
4. 10/7/2008, McCain	It is the same overall strategy. Of course, we have to do some things tactically, some of which Senator Obama is **correct** on. We have to double the size of the Afghan army. We have to have a streamlined NATO command structure. We have to do a lot of things
5. 10/3/2012, Romney	The third area, energy. Energy is critical, and the President pointed out **correctly** that production of oil and gas in the U.S. is up, but not due to his policies.
6. 10/3/2012, Romney	You say we were giving mortgages to people who weren’t qualified. That’s exactly **right**. It’s one of the reasons for the great financial calamity we had. And so Dodd-Frank **correctly** says we need to have qualified mortgages, and if you give a mortgage that’s not qualified, there are big penalties, except they didn’t ever go on to define what a qualified mortgage was
7. 10/16/2012, Romney	It’s an important one, and I think the President just said **correctly** that the buck does stop at his desk
8. 10/7/2008, Obama	And you’re **right**. There is a lot of blame to go around. But I think it’s important just to remember a little bit of history. When George Bush came into office, we had surpluses. And now we have half-a-trillion-dollar deficit annually
9. 10/15/2008, Obama	I’ll just make a quick comment about vouchers in D.C. Sen. McCain’s absolutely **right**: The D.C. school system is in terrible shape, and it has been for a very long time
10. 10/3/2012, Romney	Mr. President, you’re absolutely **right**, which is that with regards to 97 percent of the businesses are not taxed at the 35-percent tax rate, they’re taxed at a lower rate

Table 8: Excerpts from ‘A5.3 evaluation: accuracy’.

agreement with a re-stated fact (see Table 8, Examples 3, 5, 7, 9). A negative trend indicates that candidates are agreeing with such statements less frequently in recent election years, suggesting that they do not find these restatements to be correct or truthful. These combined results demonstrate an increase in language surrounding matters of factual accuracy, credibility and truth.

The trends in semantic themes of discourse also appear to mirror contemporary current events. Namely, the 'color and color patterns' domain provides evidence of semantic themes shifting in response to social events. This domain consists of colour words, including *black*, *brown*, *white* and *color*. The dominant contributor to this domain is *black* at approximately 28 percent, with the majority (57 percent) of these instances occurring during the 2020 debate series. Furthermore, all instances of the word *black* are referring to people (e.g., 'black males', 'black children', 'Black Lives' and 'Black community'). A Tau value of +0.6 indicates that these semantic representations of colour have been increasing, including a 383 percent increase from 2016 to 2020. The year 2020 was pivotal in the Black Lives Matter movement, during which time matters of race were commonly at the forefront of news and politics. This is a very large increase in this domain, led by instances of *black*, and is consistent with expectations of thematic trends within presidential debates. Although the Black Lives Matter movement existed under this moniker since 2013 and has since been growing, in 2020 this movement garnered both national and international attention to a degree it previously had not. Similarly, semantic representations of colour have been generally increasing since 2000, with 2020 accentuating this growth with a uniquely large percentage increase.

Of the four sub-domains that demonstrate nearly perfect negative correlations, the most notable is the 'toughness: strong/weak' domain with a Tau value of –0.867 and maximum percentage decrease of 59 percent between 2016 and 2020. When coupled with the negatively trending frequencies in the 'warfare, defense, and the army; weapons' sub-domain (Tau = –0.733), these trends may also reflect American contemporary history. The 'warfare, defense, and the army; weapons' sub-domain only increased between two election years – 2000 and 2004. This 79 percent increase corresponds with the first election year subsequent to the 9/11 attacks of 2001, a time during which American values prioritised national security and defence. The 2004 debates also reflect a large decrease in other prominent themes, such as 'education: in general' (–59 percent). What is normally a common concern for prospective candidates may have been over-shadowed by larger concerns of domestic safety. Moving from 2004 to 2020, the declining 'toughness: strong/weak' and 'warfare, defense, and the army; weapons' sub-domains agree with a declining prioritisation of national security, as the United States has begun to look more inward rather than outward.

5. Discussion

Although 2016 to 2020 is considered by many, perhaps conveniently so, to be an anomalous period with regards to the American presidency, the data suggest that this may not be the case for presidential debate discourse. In what may be regarded as a unique era of fake news, accusations of credibility, and a focus on people rather than policy, semantic domains in presidential debates appear to have been trending in this direction for at least two decades. Statements questioning factual accuracy have been steadily growing, reflecting an increasing concern to establish truth; pronoun usage, dominated by personal pronouns, has increased with each successive election, implying that rhetoric has been orientated around discussions of people; words of quotation and evaluations of verbal commitments have grown as well, signifying an importance of not only completing, but also making legitimate promises. It is important to note that, when considered holistically, these trends appear to transcend party lines, as the twenty-first century has thus far witnessed both Republican (twelve years) and Democratic (eight years) presidencies. Nevertheless, diachronic change is possibly the result of trends dominated by one particular party, and further analysis across party lines is necessary to determine whether or not such trends are truly shared. As demonstrated by the increasing pronoun frequency, it is possible for trends to be more strongly motivated by one political party. In either case, the data ultimately support the notion that the current state of presidential debate rhetoric is a natural progression from where the language usage of candidates was already heading. Moreover, noting prominent semantic domains that did not demonstrate statistically significant trends (e.g., ‘general ethics’, ‘business: generally’, ‘government’, ‘health and disease’, ‘the media’ and ‘green issues’) provides evidence that the semantic frequencies of many common policy subjects have generally not increased or decreased over the past twenty years. Future research may focus on how specific semantic domains do or do not differ across political parties, as well as provide a more focussed analysis on which semantic frequencies demonstrate no significant trends.

The semantic domains of presidential debates also tend to reflect contemporary American and presidential history. While causation can certainly not be implied, we see prioritisation of values (e.g., national security, education and strength) reflected by trends in relevant semantic domains. Frequency spikes in matters of warfare and defence, for instance, correspond with the 9/11 attacks of 2001, with complimentary declines in other common debate topics, such as education. Such spikes also parallel spikes in contemporary social movements, as demonstrated by ‘color and color patterns’ when considered in the context of the burgeoning Black Lives Matter movement in 2020. Closer analysis, such as the identification of key semantic domains in each election year, may provide further evidence as well.

5.1 Limitations

While presidential debate rhetoric provides a meaningful snapshot of the state of America, its social and political concerns, and the values which define the nation and the parties, the analysis is limited by the nature of the relationship between a president and their citizens. Benoit and Hansen (2001) have shown that, at least until the year 2000, journalistic questions and debate prompts were not necessarily a reflection of what presidential candidates should discuss to address public opinion. It logically follows that although a particular semantic domain may be increasing or decreasing in frequency over time, this is only a reflection of presidential candidate language during debates and should not be generalised to the language of US politics generally. Furthermore, the corpus of interest is relatively small at fewer than 250,000 words. Although a judgement of appropriate corpus size is somewhat arbitrary, corpus linguistics as a field generally investigates much larger corpora of greater than one-million words. However, in the study of specialised contexts, discourses and communities, compiling a corpus that exceeds this threshold is often problematic, if not impossible. The PD Corpus created and analysed for this study is small, but it is exhaustive in the sense that all eligible texts were included. In other words, the corpus contains all twenty-first century presidential debate discourse. To mitigate this concern, future research may include debate transcripts from prior debate years, in addition to intra-party primary debates, to expand upon more recent trends in semantic domain frequencies and provide more robust statistical support for correlation analysis.

6. Conclusion

This corpus-assisted discourse study revealed diachronic change in semantic domains that provides valuable insight into the language of the American presidency. Although the American presidency is perpetually evolving, quantitative analysis of presidential debate discourse demonstrates remarkable trends that suggest political debate discourse is a gradually shifting linguistic environment. The results of this study challenge the idea that modern political debates are an outlier in the chronology of political discourse – rather, it demonstrates a natural progression into the Trump era and beyond. The notion of gradually shifting discourse is significant, as it indicates that semantic aspects of candidates' language use may change incrementally as opposed to suddenly. Though candidates may have widely varied, unique speaking styles, the salient topics of discourse appear to demonstrate patterned usage for at least two decades. For example, these findings reveal linguistic patterns that are increasingly used by candidates, such as an orientation towards discussions of people, factual accuracy and verbal commitments. As demonstrated by both positively and negatively correlated semantic domains, the evolution of discourse in presidential debates can also evaluate conceptions of contemporary American history, language of presidential debates, and the expectations of linguistic patterns of presidential candidates. Ultimately, this corpus-aided discourse analysis provided a quantitative overview of changes in semantic domains, offering a valuable framework for identifying diachronic trends that may otherwise remain occluded.

References

Aluthman, E.S. 2018. ‘A corpus-assisted critical discourse analysis of the discursive representation of immigration in the EU referendum debate’, Arab World English Journal 9 (4), pp. 19–38.

Baker, P. 2012. ‘Acceptable bias? Using corpus linguistics methods with critical discourse analysis’, Critical Discourse Studies 9 (3), pp. 247–56.

Baker, P. and T. McEnery. 2005. ‘A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts’, Journal of Language and Politics 4 (2), pp. 197–226.

Baron, A., P. Rayson and D. Archer. 2009. ‘Word frequency and key word statistics in corpus linguistics’, Anglistik 20 (1), pp. 41–67. Available online at: https://www.research.lancs.ac.uk/portal/en/publications/word-frequency-and-key-word-statistics-in-corpus-linguistics(bf0d0d9d-10e1-4a1c-ad22-fc2635f18776).html.

Benoit, W.L. and G.J. Hansen. 2001. ‘Presidential debate questions and the public agenda’, Communication Quarterly 49 (2), pp. 130–41.

Biber, D. and E. Finegan. 1990–. ‘ARCHER: A Representative Corpus of Historical English Registers’. Available online at: https://www.alc.manchester.ac.uk/linguistics-and-english-language/research/projects/archer/.

Buckingham, L. and N. Alali. 2019. ‘Extreme parallels: a corpus-driven analysis of ISIS and far-right discourse’, Kōtuitui: New Zealand Journal of Social Sciences Online 15 (2), pp. 310–31.

Clarke, I. and J. Grieve. 2019. ‘Stylistic variation on the Donald Trump Twitter account: a linguistic analysis of tweets posted between 2009 and 2018’, PLOS ONE 14 (9), e0222062.

Davies, M. 2010–. ‘The Corpus of Historical American English (coha).’ Available online at: https://www.english-corpora.org/coha/.

El-Falaky, M.S. 2015. ‘Vote for me! A corpus linguistic analysis of American presidential debates using functional grammar’, Arts and Social Sciences Journal 6 (4), 1000123.

Fitzsimmons-Doolan, S. 2019. ‘Taxpaying, importing, enforcing: emerging discourse patterns in online newspaper comments about US immigrant education’, Journal of Corpora and Discourse Studies 2, pp. 94–116.

Frayne, C. 2019. ‘An historical analysis of species references in American English’, Corpora 14 (3), pp. 327–49.

Fredricks, G.A. and R.B. Nelsen. 2007. 'On the relationship between Spearman's Rho and Kendall's Tau for pairs of continuous random variables', Journal of Statistical Planning and Inference 137 (7), pp. 2143–50.

Golshan, T. 2016. 'Donald Trump's strange speaking style, as explained by linguists', *Vox*. Available online at: https://www.vox.com/.

Gonawela, A., J. Pal, U. Thawani, E. van der Vlugt, W. Out and P. Chandra. 2018. 'Speaking their mind: populist style and antagonistic messaging in the Tweets of Donald Trump, Narendra Modi, Nigel Farage, and Geert Wilders', Computer Supported Cooperative Work (CSCW), 27 (3–6), pp. 293–326.

Gries, StTh. 2010. 'Useful statistics for corpus linguistics' in A. Sanchez and M. Almela (eds) A Mosaic of Corpus Linguistics, pp. 271–81. Frankfurt: Peter Lang.

Gries, StTh. and M. Hilpert. 2010. 'Modeling diachronic change in the third person singular: a multifactorial, verb- and author-specific exploratory approach', English Language and Linguistics 14 (3), pp. 293–320.

Halmari, H. 2008. 'On the language of the Clinton–Dole presidential campaign debates', Journal of Language and Politics 7 (2), pp. 247–70.

Hart, R.P. 1984. 'The language of the modern presidency', Presidential Studies Quarterly 14 (2), pp. 249–64.

Hilpert, M. and StTh. Gries. 2009. 'Assessing frequency changes in multistage diachronic corpora: applications for historical corpus linguistics and the study of language acquisition', Literary and Linguistic Computing 24 (4), pp. 385–401.

Hyland, K. and F. Jiang. 2016. 'Change of attitude? A diachronic study of stance', Written Communication 33 (3), pp. 251–74.

Jacobsen, R.R. 2016. 'Reformulating the question in US presidential debates', Pragmatics and Society 7 (3), pp. 391–412.

Johansson, C. 2002. 'Pied piping and stranding from a diachronic perspective', papers from the Twenty First International Conference on English Language Research on Computerized Corpora Sydney 2000, pp. 147–62. 2002 New Frontiers of Corpus Research. (Volume 36.) Brill: Rodopi.

Lim, E.T. 2002. 'Five trends in presidential rhetoric: an analysis of rhetoric from George Washington to Bill Clinton', Presidential Studies Quarterly 32 (2), pp. 328–48.

McDonnell, A. 2020. 'Clinton stated, Trump exclaimed!' Journal of Language and Politics? 19 (1), pp. 71–88.

Mair, C. 2002. 'Three changing patterns of verb complementation in late modern English: a real-time study based on matching text corpora', English Language and Linguistics 6 (1), pp. 105–31.

Marchi, A. 2010. '"The moral in the story": a diachronic investigation of lexicalised morality in the UK press', Corpora 5 (2), pp. 161–89.

Millar, N. 2009. 'Modal verbs in TIME', International Journal of Corpus Linguistics 14 (2), pp. 191–220.

Ott, B. and G. Dickinson. 2019. The Twitter Presidency: Donald J. Trump and the Politics of White Rage. New York: Routledge.

Partington, A. 2012. 'The changing discourses on antisemitism in the UK press from 1993 to 2009', Journal of Language and Politics 11 (1), pp. 51–76.

Poole, R., A. Gnann and G. Hahn-Powell. 2019. 'Epistemic stance and the construction of knowledge in science writing: a diachronic corpus study', Journal of English for Academic Purposes 42 (19), 100784.

Proctor, K. and L. Su. 2011. 'The 1st person plural in political discourse –American politicians in interviews and in a debate', Journal of Pragmatics 43 (13), pp. 3251–66.

Rayson, P. 2002. 'USAS: UCREL Semantic Analysis System'. (Invited talk.) Tokyo, Japan: Daito Bunka University. Available online at: https://www.lancaster.ac.uk/staff/rayson/publications/tokyo2002/.

Rayson, P. 2008. 'From key words to key semantic domains', International Journal of Corpus Linguistics 13 (4), pp. 519–49.

Rissanen, M., M. Kytö, L. Kahlas-Tarkka, S. Nevanlinna, I. Taavitsainen, T. Nevalainen and H. Raumolin-Brunberg. 1991. The Helsinki Corpus of English Texts. Available online at: https://helsinkicorpus.arts.gla.ac.uk/display.py?what=index.

Romero, D.M., R.I. Swaab, B. Uzzi and A.D. Galinsky. 2015. 'Mimicry is presidential', Personality and Social Psychology Bulletin 41 (10), pp. 1311–19.

Sedensky, M. 2017. 'Trump's speaking style still flummoxes linguists'. *Hartford Courant*. Available online at: https://www.courant.com/.

Tyrkkö, J. 2016. 'Looking for rhetorical thresholds: pronoun frequencies in political speeches' in N. Minna, U. Lutzky, G. Mazzon and C. Suhr (eds) The Pragmatics and Stylistics of Identity Construction and Characterisation. Helsinki: University of Helsinki. Available online at: http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-61992.

University of California Santa Barbara. 2020. 'The American Presidency Project'. Available online at: https://www.presidency.ucsb.edu/.

Wang, Y. and H. Liu. 2018. 'Is Trump always rambling like a fourth-grade student? An analysis of stylistic features of Donald Trump's political discourse during the 2016 Election', Discourse & Society 29 (3), pp. 299–323.

Wayne, T. 2017. ‘What we talk about when we talk about and exactly like Trump’. *The New York Times*. Available online at: https://www.nytimes.com/.

Wilkinson, M. 2019. “‘Bisexual oysters”: a diachronic corpus-based critical discourse analysis of bisexual representation in *The Times* between 1957 and 2017’, Discourse & Communication 13 (2), pp. 249–67.

Wilson, A. and P. Rayson. 1993. ‘Automatic content analysis of spoken discourse’ in C. Souter and E. Atwell (eds) Corpus Based Computational Linguistics, pp. 215–26. Amsterdam: Rodopi. Also available online at: http://ucrel.lancs.ac.uk/papers/war93.txt.