<http://www.diva-portal.org>

Preprint

This is the submitted version of a paper published in *Language sciences (Oxford)*.

Citation for the original published paper (version of record):

Lundberg, J., Laitinen, M. (2020)

Twitter trolls: A linguistic profile of anti-democratic discourse

Language sciences (Oxford), 79: 1-14

<https://doi.org/10.1016/j.langsci.2019.101268>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-93507>

Twitter trolls: a corpus linguistic profile of anti-democratic discourse

Jonas Lundberg¹ & Mikko Laitinen^{1,2}

¹Linnaeus University, ²University of Eastern Finland

Declaration of Interest: The authors have no conflicts of interest, direct or indirect, that would be relevant or related to the content presented in this article.

Keywords:

Social media trolls, Twitter, anti-democratization, discourse style, personal pronouns, English as a lingua franca

Abstract:

This article focuses on anti-democratic discourse and investigates the linguistic profile of Twitter trolls. The troll data consist of some 3.5 million messages in English obtained through Twitter in late 2018. These data originate from potentially state-backed information operations aimed at sowing discord in Western societies. The baseline data, against which the troll data are compared, contain circa 4.4 million messages in English drawn from the Nordic Tweet Stream corpus. A machine learning application that enables us to select genuine personal messages in this corpus is used to prune the data. The empirical part investigates frequency-based characteristics of the two datasets. We utilize a set of automatically-extracted word-list information and the observed frequencies of personal pronouns. Our empirical findings show considerable quantitative differences so that the troll data are shorter, make use of a smaller number of lexical types and tokens, and resemble more formal registers, while the personal messages are more spoken-like. The results could be used to improve automated detection systems whose purpose is to identify troll accounts.

Twitter trolls: a linguistic profile of anti-democratic discourse

Jonas Lundberg¹ & Mikko Laitinen²

¹Linnaeus University, ²University of Eastern Finland

1. Introduction

This article investigates anti-democratic discourse in one social media application. It focuses on Twitter, which is a microblogging platform that enables users to post short messages with a maximum of 280 characters and to include links and multimedia content in the messages.

Twitter users can also participate in message exchanges and re-tweet what others have written. One special characteristic of all social media applications is that they not only contain messages from real genuine people but also material created and distributed by software robots, or bots. While the bot population consists of a heterogeneous set of account types (Morstatter et al. 2016), the common denominator is that they are non-personal and often automated accounts that post content onto online social networks. One example of an automated bot account is a weather bot that posts weather data at regular intervals, which makes it a harmless account type whose presence only skews the results of academic research using social media data. However, another account type that is far from harmless consists of non-personal online trolls. According to Coles & West (2016: 234), trolls are malicious accounts that are employed for various purposes, such as posting misinformation and spreading fabricated news online. One underlying reason for the presence of trolls is that they are typically associated with anti-democratic forces. In recent news media reports, troll accounts have been said to be capable of swinging close democratic elections and thus influencing popular opinion and breeding discord in democratic societies. In recent times, trolls have also been associated with a larger phenomenon of misusing digital information in the so-called “Blitzkrieg for the digital age” (Guardian 2018).

It is beyond the scope of this article to discuss the typology of bots. Rather, we focus on the linguistic profiling of troll accounts and, using frequency-based information, we compare some of the ways in which troll messages differ from material that represents real humans who use Twitter. Our definition is a simplification since troll accounts are, of course, maintained by real people, but as we show below, the datasets used here differ considerably in terms of how personal or impersonal they are.

The empirical material consists of two sets of data. On the one hand, as for the troll data, we utilize some 3.5 million messages in English obtained freely through Twitter in late 2018 (see https://about.twitter.com/en_us/values/elections-integrity.html#data). According to Twitter, these messages were released as part of their attempt to tackle false information and fake news, and they are from publicly available archives that Twitter believes have “resulted from potentially state-backed information operations” inflicted on their service. That is to say, they represent anti-democratic discourse.

On the other hand, we have used circa 4.4 million messages in English drawn from the Nordic Tweet Stream (NTS) corpus. This corpus is a multilingual monitor corpus of tweets and their metadata from five Nordic countries (Laitinen et al. 2018). We use a machine learning application that enables us to zoom into genuine personal accounts in the NTS. This corpus provides the baseline data against which the troll data are compared.

By concentrating on the linguistic profiling of Twitter trolls, this article charts a territory that has thus far received little attention in English corpus linguistics and in the study of the democratization of media discourse. First, even though Twitter data are from one social media application, the empirical results can be useful elsewhere. For instance, our data represent English as a lingua franca (ELF) use obtained from settings in which English is used as a non-native resource (for the definition of ELF, see Mauranen 2018). Initial forensic analyses of Twitter trolls have concluded that the syntactic constructions clearly suggest “a distinct regional origin where English is a non-native language” (Boyd et al. 2018: 5). We wish to expand this by focusing on comparing troll data with non-native use. Our observations, therefore, introduce a much-needed ELF perspective into the debate on democratization in language use. The phenomena of democratization and (anti-)democratic discourse are indisputably global, and undoubtedly require collaboration between first- and second-language scholars, including areas such as ELF, as is the case here, and such aspects as learner language and language acquisition. More empirical evidence from ELF is needed to better understand the global dimensions of the democratization of discourse.

Second, our work is heavily interdisciplinary, combining sociolinguistics with computer science and machine learning. The baseline data, for example, is part of a real-time data stream, and the machine learning application that identifies automated and non-personal messages is the result of our interdisciplinary collaboration (Lundberg, Nordqvist & Laitinen 2019). This application increases data accuracy in the real-time data streams used in sociolinguistics and makes it possible to focus on genuine personal accounts (see Section 3).

The empirical part of this study investigates the linguistic profiles of the two datasets. With respect to the research question, we aim to study the extent to which the datasets differ quantitatively from each other. We utilize a set of automatically-extracted word-list information and observed frequencies of personal pronouns. As these quantitative features appear in high frequencies, their use is most likely below the level of linguistic awareness (Labov 2007; Tagliamonte 2012: 57–58). In consequence, this type of information obtained through corpus linguistic profiling could also be used in the future to improve automated detection systems whose purpose is to identify troll accounts.

In what follows, we first present an overview of democratization as discourse-pragmatic practice and discuss how our approach is related to a set of previous studies of this topic. Section 3 offers an overview of the two datasets used. In Section 4, we look first into the quantitative profiles and basic linguistic characteristics of the data, using general frequency-based and word list information. It is clear that the only known parameters of the datasets are based on external

criteria that suggest whether the data consist of troll messages or genuine personal messages, and this section aims at establishing whether the two datasets are indeed different from each other. Section 4 therefore constitutes the foundation for our analysis of democratization markers. In Section 5, we then focus on personal pronouns as overt markers of power (a)symmetry, while Section 6 offers our conclusions and thoughts for further studies.

2. Background

Our approach builds on previous corpus-based research into the democratization of discourse, which typically sets out by distinguishing a dual relationship between language use and the socio-cultural contexts within which language use occurs (Leech et al. 2009: 259; Farrelly & Seoane 2012). One part of this duality consists of recognizing how changing societal norms in personal relations are reflected through language in use. Previous research in this line of thought has investigated such matters as frequency shifts in modal auxiliaries and has observed, for example, a decreasing use of deontic *must* and its gradual replacement with less face-threatening alternatives, such as *HAVE to* and *NEED to* (Myhill 1995). The other part of the duality means that language use shapes and causes socio-cultural change. With regard to the latter, we see democratization as a macro-level phenomenon in which online troll accounts investigated here operate as an anti-democratic force. The sole purpose of their existence is that they should be part of state-backed information operations that aim at shifting power balances within Western societies, primarily in North America and Europe. In other words, linguistic resources are being used to forge socio-cultural and political change. Without going into excessive detail, it has been suggested that malicious accounts can be used to mislead and manipulate social media discourse (Ferrara et al. 2016). In addition, such accounts may artificially inflate information and opinions.

A quick glance at our troll data shows that they contain a heterogeneous set of messages. They comprise messages describing personal events (1)–(3), and neutral messages (4)–(6) in the form of aphorisms and sports news, but also information that is fabricated and whose truth value does not satisfy the journalistic requirement, as in (7)–(10). Note that all of the material is reproduced here in the format that it appears in in the dataset released by Twitter, and the messages can be found in the elections integrity data in the URL mentioned in the introduction. (NB. Unless specifically marked, the examples are from the troll dataset).

(1) I'm gonna start my night it just a few minutes Yall

(2) It's funny how I'm good at giving advice to others, but when it comes to helping myself, I don't know what to do

- (3) I Love Music but i hate what you call music....I'm looking at you "music concrete"
- (4) I need to live my own life, follow my own path. I wish for things to change, and that's the problem. –Merida (Brave)
- (5) Saints select DT Tyeler Davison of Fresno State with second fifth-round pick
- (6) Watch highlights from LSU's 48-20 win over WKU
- (7) You want your Country back?. He knows how to close the borders and keep Muslims out
- (8) I believe that Jordan does everything right about ISIS
- (9) Abdul Malik Abdul, Arizona, faces a terrorism charge over providing guns and support to two followers
- (10) RT : Obama: "Islam has a proud tradition of tolerance" Is it tolerant to behead then?

As for the personal accounts in the NTS corpus, we cannot claim that they are exactly the opposite, i.e., completely democratic and harmonious. On the contrary, they contain daily online interaction in all of its diverse forms, including arguments, views that promote opposition politics, attempts to overturn political hegemony, or simply incoherent noise. Despite this diversity, we assume that they represent interaction in which the primary aim is to maintain interpersonal relations and engage in outward communication, in which a message can be accessed by almost anyone and where the messages are aimed at a wide readership (Honey & Herring 2009). The first part of the empirical results, which presents a set of automatically-extracted frequency data, aims at uncovering the extent to which language use that shapes and causes socio-cultural and political change is different from normal outward communication in the social media.

On the other hand, we see democratization as a micro-level phenomenon in which people use linguistic resources in response to the socio-cultural events around them. We examine the use of overt sociolinguistic markers that are typically associated in previous studies with the democratization of discourse. We therefore approach democratization from the perspective of “speakers’ tendency to avoid unequal and face threatening modes of interaction” and the propensity to use linguistic items that are more often used in speech-like styles in written texts (Farrelly & Seoane 2012: 393). The former has sometimes been called democratization proper, while the latter refers to the colloquialization of style (Mair 1997).

First, with regard to the approach whereby democratization is reflected in the discourse-pragmatic processes, third-person pronouns are considered to be one of the overt markers of power asymmetry. That is, studies that have looked into democratization proper have observed

aspects such as the decline from the early modern period onwards of generic *he* or the presence of singular *they* (Laitinen 2008). With regard to more recent changes, various studies have reported substantial shifts in the frequencies of third-person pronouns. For instance, Romaine (2001) has observed that men are referred to three times more often than women in Standard English corpus data. In contrast, Baker (2009) shows a slight increase in the use of feminine third-person pronouns and an incipient decrease in masculine pronouns in the Brown family of corpora, which represent Standard English in the twentieth and early twenty-first centuries. He adds that although the feminine pronouns occur collectively less frequently than their male counterparts, “male bias may be decreasing” (2009: 328).

Second, pronoun use is closely associated with colloquialization. As an illustration, pronoun distribution characterizes prototypical discourse styles, so that first- and second-person pronouns are more frequently used in involved discourse situations in which the interlocutors aim at maintaining interaction, but speakers tend to use third-person pronouns in narrative contexts in which the objective is to transmit information (Biber 1988: 101–109). The more involved a situation is, as opposed to being merely informational, the more spoken-like it is.

It has for a long time been obvious that digital communication and language use in social networks have blurred the line between spoken and written communication because the mode is written, but linguistic resources are being used for interactive purposes (Knight et al. 2014). The new component here is the fact that we can assume that genuine personal accounts and troll data are different because they exist for different purposes, and once we have presented our dataset, we will focus more sharply on the question of exactly *how* different they are.

3. Material

Our baseline data, the NTS corpus, has been storing material since November 2016. The data collection is automatic and uses HBC (<https://github.com/twitter/hbc>) as the downloading mechanism. The data capture builds on the geolocation information, for which we first specify the geographic region covering the five Nordic countries, and a second filtering process is then applied to select only tweets tagged with a Nordic country code (DK, FI, IS, NO or SE). This second filtering is necessary in order to exclude tweets from neighboring countries (e.g., Germany and Russia) located within the chosen geographic boundary (see Laitinen et al. 2018).

Before we can utilize these data as our baseline, we apply a third filtering process that removes automatically-generated bot messages. According to Morstatter et al. (2016), bot refers to a heterogeneous set of account types that are non-personal and that post automated messages, and their presence can seriously skew sampling and result in inaccuracies when used as

empirical evidence. Previous sociolinguistic studies have relied on a range of methods when it comes to handling bot accounts. Huang et al. (2016), for instance, recognize their presence but include them in their results (which is also true for Laitinen et al. 2017). Coats (2017) utilizes a method in which material from certain types of devices is excluded. In Laitinen & Lundberg (in press), we have utilized a language-specific algorithm to increase data accuracy and remove automatically-generated tweets (AGT).

To clean the data in the present study, we have utilized a supervised machine-learning algorithm. This algorithm was developed by Lundberg et al. (2019) and is a language-independent approach for detecting AGTs. We define an AGT as a tweet in which all or parts of the natural language content are generated automatically by a bot or other type of program. The system is language independent since the actual Twitter text is not used as an input feature in the classifier. Rather, the algorithm classifies each tweet using only select attributes in the metadata, which are available for each tweet. This feature makes our approach simple and light so that it can handle high-velocity data. In our experiments, the best classifier (J48-HP) had an error rate of 1.84%, with precision 0.998, and recall of 0.920. This application enables us to focus solely on human-generated tweets (HGT). Note that the application only identifies social bot messages, and it is possible that the baseline data contain messages from troll accounts.

The baseline dataset NTS-HGT contains the English-language tweets sent during 2017–2018 from within the Nordic region, a total of 4,360,261 messages. With regard to whether the data represent English use in the region and not messages from tourists and visitors, previous studies that use Twitter data have suggested that a great majority of messages in one location such as a single country are in fact from residents of that location (Lamanna et al. 2018). We therefore assume that our baseline dataset is reliable, given the general limitations of Twitter data in general, while certain demographic groups, such as younger generations (male in particular) and public figures, may well be over-represented. The NTS-HGT data represent genuine personal messages.

In the case of the anti-democratic troll accounts, we use two sets of material that have been made available for data-intensive research on the understanding that the research activities comply with the Twitter Developer Agreement and Policy. The datasets consist of material from 3,613 accounts from *the Internet Research Agency*, which in popular accounts is held to be a Russian government-owned troll factory that engages in cyber operations worldwide. Whether this is indeed true we do not know with complete certainty but have to rely on indirect evidence. As an illustration, CNN reported in spring 2019 that the agency had “engaged in a years-long campaign to sow discord in the U.S....by creating and maintaining fake social media

personas and activist organizations designed to look like they were run by real Americans” (CNN 2019). In addition, we use information from 770 accounts that are believed to originate from Iran. According to the information available, both of these datasets comprise public messages and media (such as images and videos) from accounts that are allegedly part of information operations aimed at spreading false news and sowing discord in Western societies.

Altogether, these troll data contain 3,538,681 messages in English. According to Twitter, the dataset is thought to be comprehensive as it contains all non-deleted material obtained from suspicious accounts. The material that had been deleted by account holders represents less than 1% of the total activity. Note that the accounts in these data are no longer active but were suspended by Twitter prior to when the raw data were made available.

We downloaded the troll data in early 2019 and pruned them to include only the textual material, while excluding all the multimedia content. In addition, simple start and end tags were inserted around URL-addresses, hashtags, and mentions/re-tweets, thus enabling us to focus solely on the content keyed in by the users. After this, the baseline data and the troll data were automatically parts-of-speech tagged using spaCy and the Penn treebank tagset (<https://github.com/explosion/spaCy>).

In the empirical part, one dataset represents genuine human social media interaction (NTS-HGT), while the other, according to Twitter, was part of information operations aimed at spreading false news online (Troll data). The datasets are comparable, and even though they originate from different areas, they represent written ELF use. We have used the NTS data in various ELF studies (Laitinen et al. 2017; Laitinen & Lundberg in press; Alissandrakis et al. in press). With respect to the troll data, we know very little about the real identities of the authors. We assume the majority of the authors to be non-native English speakers. A noteworthy feature of the Russian trolls was reported in the news in spring 2019. According to respectable news sources, many high-ranking individuals and credible news outlets in the USA used and quoted the tweets now included in the troll dataset during the presidential campaign in 2016 without knowing that they originated from troll factories (CNN 2019). The following section presents the basic metrics of the two datasets.

4. Results

4.1. Frequency-based profile of the datasets

The first two sections present a detailed analysis of the frequency-based patterns in the datasets and serve as a precursor to the following empirical section, which deals with pronoun use. The

objective is, first, to utilize quantitative evidence and to explore the extent to which the two datasets are similar, and then to focus more closely on a linguistic category.

Table 1 presents the raw frequencies of the datasets. The number of tokens in the NTS-HGT is slightly over 53.5 million, and it includes all of the HGT messages in English dating from 2017–2018. The respective figure in the troll data is circa 39.6 million words.

Table 1. Basic word-list frequencies in the datasets

	NTS-HGT 2017–2018	Troll data
Number of tweets	4,360,261	3,538,681
Tokens	53,541,861	39,635,886
Types	551,037	302,516
Type/token (<8 tokens)	0.97	0.97
Type/token (9–15 tokens)	0.94	0.95
Type/token (>16 tokens)	0.88	0.90
Prepositions (per 100,000)	12,614	12,614
Average complexity	1.54	1.41
Average tokens/message	12.28	11.20
Average types/message	11.13	10.48

The datasets offer insight into ways in which genuine personal tweets differ from or are similar to troll accounts. First, since the average length of the messages differs slightly, we have divided the type-token ratios for three length categories. For the shortest messages (<8 tokens), the ratio is identical (0.97), and remains so also for the longer messages, with only slight differences. Even though the troll data is primarily limited to political campaigns and therefore most likely restricted to a smaller set of potential topics, it is not reflected in lexical richness. This finding is similar to what has been found in previous studies that investigate automated bots. For instance, Ferrara et al. (2016) observe that the distinction between humanlike and bot-like behavior has become increasingly fuzzier, since bots (and trolls) can engage in complex social interactions by generating suitable content and identifying relevant keywords.

Second, another measure Table 1 is that of message complexity. We estimate complexity using the number of prepositions in the datasets and per message. The assumption is that the presence of prepositions indicates that more complex nominal phrases are present, and previous experimental studies have used this measure as an index for detecting deception in spoken language. Thus, Briscoe et al. (2014: 1438) observe that higher complexity tends to result in the lower likelihood of a speaker being identified as deceptive. However, it should be noted that the threshold values that Briscoe et al. use are substantially higher than the frequencies in Table 1. The normalized frequencies of prepositions are identical (12,614), and

there is a slight difference per message (1.54 vs. 1.41). Overall, the tweet complexity of <2 prepositions per message would be considered to be low in terms of sentence complexity. Given the short message length in general in the Twitter data, the difference is considerable but should most likely not be used independent of the other quantitative measurements.

Third, the main quantitative pattern that is visible in Table 1 is the average message length. The NTS-HGT tweets are consistently longer in terms of both average number of tokens and types. These differences are statistically highly significant when measured using a two-tailed t-test for independent samples (for tokens $t=93.98$, $p<0.0001$; for types $t=65.91$, $p<0.0001$).

Figure 1 visualizes the differences using 1 million randomly selected messages from the data. For the sake of brevity, it only shows the number of tokens per dataset (NTS-HGT vs. Troll), but it should be noted that the pattern is almost identical for type-related information. The visualization illustrates that while the mean values are close to each other, the NTS-HGT messages are consistently longer than the troll messages, and this difference is most clearly visible in the spread, which in the HGTs is considerably larger than in the troll dataset.

Figure 1. The number of tokens in 1 million randomly selected messages

Moreover, the difference is particularly noticeable in the upper quartiles and the upper ranges of the message length in the HGT data.

Figure 1 summarizes the frequency patterns in a large subset of the data, but a more illustrative method would be to select a smaller randomly selected sample that would show the spread of items. In Figure 2, we have selected a total of 1,000 messages to illustrate the differences in the number of distinct lexical types keyed in by the account holders. The figure shows the NTS-HGT accounts in red and the troll data in green.

Figure 2. The number of types in the NTS-HGT (red) data and the Troll (green) data

Figure 2 makes the differences even clearer, as the personal messages have a broader spread of type numbers, while the troll data clearly have the densest concentration of between 5–20 types. To be add to the visual information, if we calculate mean lengths and standard deviation in 100 iterations, the 70% of the troll data are between 5–15 types, while the spread in the genuine human tweets is between 3–20. Moreover, this visualization suggests that it is unlikely that an anti-democratic troll message would have more than 25 types, but a sizeable set of NTS-HGT messages contain a much richer selection of lexis.

Our findings here from Twitter are partly orthogonal to what has been suggested in computer-mediated communication (e-mail) interaction. In experimental settings, Burgoon et al. (2003: 92) find that “deceivers were more likely to use longer messages but with less diversity and complexity” than those who were honest. The term “deceivers” is comparable to the troll messages here. However, Burgoon et al.’s findings also show that in spoken face-to-face interviews deceivers tended to use shorter messages in terms of the number of tokens, but with a larger stock of lexical types.

4.2. Word-lists and 2-grams

Our next focus is on the most frequent types and their frequencies in the data. Word-list comparisons have been commonly used in previous studies as a similarity measure. Baker (2009), for example, compares the rankings of the 20 most frequent types as an indicator of ongoing change in the four Brown corpora of Standard American and British English (using BLOB-1931, LOB, FLOB and BE06 corpora). Similarly, Oakes (2009) explores the rankings of the most frequent lexical items to measure the correlation between spoken and written Standard English and finds that corpus (dis-)similarity is probably reflected in the rankings of high frequency words.

Table 2 below presents the thirty most frequent types (case insensitive) in the data, and shows that the differences in the lexical profiles of the NTS-HGT and the troll accounts are noticeable in the types and their proportional frequencies. Note that the frequencies consist of raw frequencies from two datasets of different size.

Table 2. 30 most frequent lexical types in the datasets

NTS-HGT	Freq.	Prop	Troll	Freq.	Prop.
the	1,768,461	3.21	rt	1,291,828	3.21
i	1,462,702	2.66	the	1,046,176	2.60
to	1,272,914	2.31	to	982,638	2.44
a	1,127,376	2.05	in	707,518	1.76
and	1,016,408	1.85	a	617,123	1.53
you / u	882,756	1.60	of	599,686	1.49
in	870,873	1.58	s	546,098	1.36
of	774,663	1.41	i	467,610	1.16
is	734,549	1.34	you / u	459,145	1.14
it	688,956	1.25	is	456,469	1.13
for	589,202	1.07	and	435,081	1.08
s	544,395	0.99	for	432,723	1.07
my	499,120	0.91	on	322,858	0.80
that	458,547	0.83	it	300,857	0.75
this	454,659	0.83	t	248,616	0.62
on	425,709	0.77	that	232,927	0.58
with	384,366	0.70	with	220,759	0.54
t	352,265	0.64	trump	200,010	0.50
so	339,718	0.62	are	187,442	0.47
have	311,405	0.57	at	184,814	0.46
at	287,299	0.52	this	184,394	0.46
be	285,277	0.52	be	180,929	0.45
are	283,285	0.51	we	159,325	0.40

but	271,977	0.49	not	157,320	0.39
me	269,780	0.49	my	142,339	0.35
we	267,121	0.49	what	133,813	0.33
not	251,628	0.46	your	133,649	0.33
just	249,408	0.45	by	125,628	0.31
m	241,377	0.44	from	124,381	0.31
can	224,336	0.41	have	124,146	0.31

The two lists are identical for 24 of the 30 items, and the bold font indicates those that are present in the NTS-HGT but not in the troll data, and *vice versa*. Some of differences are more obvious than others, such as *rt* (re-tweet) and *trump*, which are both data and context dependent. Re-tweets have been observed to be a highly frequent item in bots (Gupta et al. 2013), but *trump* is clearly context dependent, referring to Donald Trump. Other noteworthy differences are the presence of *so*, which is frequently used as an intensifier, and the tokens related to first-person discourse (*me* and '*m*') in the NTS-HGT, as these items are indicative of spoken discourse style.

An item that appears in the troll data but not in the NTS-HGT list is *by*, which suggests one substantial difference between the datasets. The troll data seem to prefer passive constructions more than do the real human messages. This small detail suggests that troll data are more news-like, because passives in general are common in news and in text types in which the share of finite verbs is lowest (Biber et al. 1999: 476). As the evidence below will show, this also appears to be the case in the troll data (see Table 4 below). In addition, the sizeable set of the passives in the troll data seem to appear in the long form, where the agent is expressed by a *by*-phrase. It is widely held that the *by*-phrase is optional, and the majority of passive sentences in English occur without it (Quirk et al. 1985: 164–165; Biber et al. 1999: 476–477, 940–943). Its omission occurs in contexts where an overt agent is irrelevant, redundant, or unknown, but it is often used in those cases where there is a need to highlight the agent, which in turn affects the participant chosen as the subject. The higher use of *by* in Table 2 may be indicative of the greater likelihood of contextually relevant *by*-agents in the troll data, as in (11) below.

- (11) Suspect shot **by** homeowner in eastern New Orleans home burglary

The wordlist and the rank order in Table 2 should be seen solely as the starting point, but they nevertheless reveal other notable patterns. For instance, the NTS-HGT data display a higher preference for interpersonal items, such as the personal pronouns (*I*, *you*, *my*, *me*), which constitute 5.67% of the data, while their proportional frequency in the troll data is only 2.63%.

The difference is substantial, which can be illustrated by the fact that every 20th token in the personal messages is any one of these high-frequency personal pronouns, while in the troll data they appear in only every 50th token. This topic will be dealt with in the following section.

Figure 3 visualizes the normalized frequencies of the 30 types. It illustrates that the datasets differ with regard to the consistently lower normalized shares of the most frequent types in the troll data. The normalized frequency is per 1 million words.

Figure 3. Frequency profiles of the fifty most frequent types in the datasets

The differences are independent of the total number of tokens in the datasets and illustrate a noteworthy difference in that natural human interaction differs from the deceptive language use in the troll data. The normalized frequencies in the troll data are lower and, with the exception of three types (*to*, *and*, *for*), the pattern is consistent.

Together with the quantitative evidence presented here, this observation highlights the differences in the linguistic profiles of the two datasets. Anti-democratic troll messages are consistently shorter, use a lower number of types and tokens, and the proportional share of the most frequent types is lower than the corresponding figure in the genuine personal messages.

We next examine 2-grams, with Table 3 below listing the 30 most frequent items. We have excluded 2-grams of function words, such as preposition + article and conjunction + article pairs. To establish any differences, it is worth investigating the list both quantitatively and qualitatively. With regard to quantitative differences, the NTS-HGT data show consistently

higher normalized and proportional frequencies than the troll data. The finding is similar to what is presented in Table 2 above and suggests that this pattern clearly distinguishes trolls from personal accounts.

With respect to qualitative differences, a range of differences emerge. The first thing to notice is the set of context-dependent items, such as *trump's*, and *donald trump*. Indeed, in the top-100 2-grams references to *Trump* outnumber the mentions of any other personal names. For instance, the ratio of *Trump:Clinton* is 5:1 in the troll data. Moreover, the datasets also differ with regard to two important features. On the one hand, the troll data also contains more 2-grams with a negative particle, and there are seven pairs in which one of the items is *not/n't* (indicated in bold). On the other hand, the NTS-HGT data contains substantially more first-person bi-grams (indicated in italics). There seems to be a greater propensity for self-references in the genuine personal data, which are much less frequent in the anti-democratic data. In addition, the troll data contains slightly more negative utterances than does the NTS-HGT material.

Table 3. 30 most frequent 2-grams in the two datasets

NTS-HGT	Norm. freq.	prop	Troll	Norm. freq.	prop
<i>i_m</i>	3,772	0.38	don_t	2,199	0.22
it_s	3,410	0.34	it_s	2,011	0.20
don_t	2,166	0.22	<i>i_m</i>	<i>1,801</i>	<i>0.18</i>
to_be	1,392	0.14	<i>rt_i</i>	<i>1,771</i>	<i>0.18</i>
this_is	1,313	0.13	to_be	1,329	0.13
have_a	1,289	0.13	rt_the	1,253	0.13
can_t	1,260	0.13	is_a	1,110	0.11
<i>and_i</i>	<i>1,218</i>	<i>0.12</i>	can_t	973	0.10
thank_you	1,212	0.12	u_s	917	0.09
is_a	1,069	0.11	if_you	881	0.09
<i>i_have</i>	<i>1,025</i>	<i>0.10</i>	is_the	844	0.08
<i>i_love</i>	<i>1,016</i>	<i>0.10</i>	you_re	831	0.08
<i>i_am</i>	<i>1,010</i>	<i>0.10</i>	this_is	799	0.08
that_s	973	0.10	that_s	725	0.07
<i>i_just</i>	<i>964</i>	<i>0.10</i>	trump_s	691	0.07
<i>i_can</i>	<i>959</i>	<i>0.10</i>	you_can	621	0.06
you_re	922	0.09	you_are	557	0.06
<i>i_ve</i>	<i>918</i>	<i>0.09</i>	want_to	542	0.05
if_you	917	0.09	it_is	529	0.05
is_the	902	0.09	donald_trump	529	0.05
<i>i_was</i>	<i>853</i>	<i>0.09</i>	doesn_t	528	0.05
you_are	790	0.08	rt_this	523	0.05
it_is	751	0.08	i_don	496	<i>0.05</i>
want_to	729	0.07	the_world	491	0.05

<i>i_don</i>	725	0.07	will_be	471	0.05
so_much	709	0.07	when_you	469	0.05
it_was	702	0.07	is_not	460	0.05
one_of	688	0.07	won_t	452	0.05
the_best	684	0.07	rt_you	442	0.04
to_see	682	0.07	we_are	433	0.04

Another noticeable feature can be deduced from the token-level list: the high frequency of bi-grams that contain the marker for re-tweets (rt). The higher propensity of re-tweets in social bots has also been observed by Ferrara et al. (2016).

With respect to the last quantitative profile in this section, Table 4 presents the frequencies of the major parts-of-speech (POS) classes in the datasets. It shows the normalized frequencies of 13 POS categories (per 1 million words) and illustrates the size of the difference, comparing the troll data with the baseline, while the right-most column shows the log-likelihood scores (calculated using the LL calculator at <http://ucrel.lancs.ac.uk/llwizard.html>). The statistical significances were of course calculated using the raw observations.

Table 4. Frequencies of the major POS categories in the datasets (***) = $p < 0.001$

		NTS-HGT	Troll	% in Troll	ll
1.	Noun	282020	401181	142	***
2.	Verb	197103	183720	93	***
3.	Punctuation	139777	155600	111	***
4.	Pronoun	103333	69668	67	***
5.	Preposition	103139	103044	100	n-sig.
6.	Adverb	83437	54987	66	***
7.	Article	77146	57274	74	***
8.	Adjective	76959	61985	81	***
9.	Conjunction	28790	17736	62	***
10.	TO-inf	15426	16093	104	***
11.	Interjection	9896	3958	40	***
12.	Possessive	2581	8700	337	***
13.	There_ex	1386	1099	79	***

The POS categories are nearly identical in terms of punctuation (111%), prepositions (100%), and *to*-infinitives (104%). Note that the complexity measurement used in the previous section took into account the number of prepositions per message rather than per dataset.

With regard to the differences, there appears to be a strong preference for more frequent noun use in the troll data; these contain 42% more nouns than do the personal accounts in the NTS. This is also illustrated by the fact that Table 2 above, which lists the most frequent lexical types, contains a range of nouns such as *trump*, *people*, *man*, *police*, *obama*, *love*, *day*, *life*,

president, video, workout, isis, and world, while the NTS list includes only function words. The higher frequency of nouns in the troll data is also reflected in the substantially higher proportion of occurrences of the synthetic possessive 's, whose share is 337% higher than that found in the NTS.

However, when attention is paid to whether the higher use of nouns in the troll messages is reflected in the article use, the results are the opposite: there are fewer articles in the troll data than in the NTS-HGT material. We return to the explanations in Section 6, but there are various possible reasons for these differences. On the one hand, variable use of articles is an ELF-specific feature; the omission of articles is common in ELF, where they may also be used superfluously (Cogo & Dewey 2006; Mauranen 2012: 124–125). The datasets obviously differ with regard to geographic origin, since the NTS messages originate from the Nordic region while the troll data come from Russia and Iran, but it must be emphasized that all of the data are from non-native-English settings. On the other hand, the troll data contain tweets in which news headlines are repeated, which could explain the lower frequency of articles. These differences are substantial and need to be studied further.

Human accounts, by contrast, contain a greater number of items in all of the other categories. Notable differences appear in relation to pronouns, adverbs, and interjections. Interjections, in particular, are rarely used in the troll data, and they are typically seen to index an informal and interactive style of communication, making the messages more human-like by marking speaker stance and highlighting interpersonal function in communication. This can be illustrated by (12), below.

(12) Ahh_UH yeah_UH ,_, I_PRP also_RB have_VBP some_DT domains_NNS :)_:
and_CC probably_RB quite_PDT a_DT few_JJ other_JJ things_NNS as_RB well_RB
!_. (NTS-HGT)

Overall, the higher proportion of nouns may well point to the troll data being more formal, but further evidence is required to prove this assumption.

These first two sections have presented a macro-level frequency-based view of the two datasets. There are clear quantitative differences between the datasets, and the troll messages are typically shorter and contain a smaller number of lexical types. Moreover, genuine personal messages are more colloquial and spoken-like, whereas features that commonly appear in more formal registers, such as news, are over-represented in the troll data. Lastly, the use of pronouns

is clearly different, as has been shown here. Hence, the following section will turn its focus to pronoun use.

5. Observations of pronoun frequencies

As shown in the previous section, pronoun frequencies and use clearly distinguishes the personal accounts from the troll data. The frequencies show that the share of pronouns in the troll data is almost one-third smaller than in the NTS-HGT data (Table 4). Moreover, as the list of the most frequent types (Table 2) indicates, there are remarkable differences, to the extent that the first-person *I* is almost three times more frequent in the personal messages than in the troll data. Similar tendencies may also be observed in the second person (*you/u*) and in the plural *we*. In addition to these quantitative differences, pronoun use is central to the democratization proper discussed in previous studies (Section 2).

In consequence, this section examines the ways in which the datasets differ from each other with regard to pronoun frequencies. Figure 4 visualizes the share of all of the pronouns in the two datasets. It reveals that the proportional share in the NTS-HGT is 10.5% of the major POS categories, while the figure in the troll data is, at 7.1%, significantly lower ($p < 0.0001$, log. likelihood value 3201.20).

Figure 4. Relative proportion of all pronouns in the datasets

If we use these broad tendencies to extrapolate differences in discourse styles, it is obvious that the NTS-HGT data are closer than the troll messages to conversational styles. Genuine personal messages are more involved and interactive than the anti-democratic troll messages. This observation can be contrasted with the pronoun frequencies obtained from multiple registers in large-scale corpus descriptions. Biber et al. (1999: 333) observe that pronouns in general are most common in conversation and that their frequencies distinguish different registers in that more formal contexts contain fewer pronouns than do interactive conversations. Our findings with respect to the pronoun frequencies in the NTS-HGT data are also similar to those of Knight et al. (2014), in which they used a very small, purpose-built corpus of tweets collected from popular public accounts with more than 1,000 followers discussing various topics.

This finding also supports the quantitative evidence of substantial differences between the datasets. The two previous sections, above, have demonstrated that troll messages employ a smaller set of tokens than do personal accounts, while the troll messages also resemble the news as a result of their higher proportion of *by*-agents in passives. Finally, troll messages are richer in nouns but contain far fewer of the interjections that characterize natural emotions and express users' stances on particular utterances.

Thus far we have looked at all of the pronouns collectively, but in the following we attempt to analyze the figures obtained for the various pronoun categories. Personal pronouns are grouped into six broad classes, as shown in Table 5. These consist of the first-person singular and plural pronouns in the nominative, possessive and objective/reflexive forms. The third category consists of the second-person *you*. The third-person pronouns include the masculine and feminine forms and their plurals, each with three categories. We exclude impersonal *it/its*.

Table 5. Personal pronoun categories

	1st sg.	1st pl.	2nd sg/pl.	3rd m.	3rd f.	3rd pl.
Nom.	<i>I / I'</i>	<i>we/we'</i>	<i>you</i>	<i>he/he'</i>	<i>she/she'</i>	<i>they/they'</i>
Poss.	<i>my/mine</i>	<i>our/ours</i>	<i>your/yours</i>	<i>his</i>	<i>her/hers</i>	<i>their/theirs</i>
Obj./refl.	<i>me/myself</i>	<i>us/ourselves</i>	<i>-self/selves</i>	<i>him(self)</i>	<i>herself</i>	<i>them(selves)</i>

Note that, as a result of the case differences between the paradigms, the categories are not identical, but the quantitative differences nevertheless serve to illustrate how these items are used in the datasets. The searches were case insensitive in that the first-person includes both *I*

and *i*. In addition, as in Baker (2009), we included forms with apostrophic *he'*, *they'*, etc., where the pronoun is followed by a contracted verb.

Figure 5 visualizes the pronoun categories and demonstrates how the normalized frequencies vary between the datasets, with the lines indicating the differences. As might be expected on the basis of the evidence in Section 4 above, the frequencies of the pronoun categories are substantially higher in the NTS-HGT data than in the trolls. However, the differences have not been caused by the larger frequencies in all of the categories in the personal message data, but rather there are only minor differences between most of the pronouns. The largest differences actually occur in only five categories: the first-person pronouns (*I*, *my_mine*, *myself*), second-person *you*, and the first-person plural *we*.

Figure 5. Pronoun categories and the frequency differences between the datasets

In sum, the interactive and the more colloquial style, measured in terms of the higher share of personal pronouns in the NTS-HGT, is brought about by only a small set of types. In addition, there are pronoun categories in which the situation is the opposite, where the frequencies are actually higher in the troll data. These are primarily third-person singular forms, but the differences are not statistically significant, except in the case of one pronoun type. The troll data contains more examples of the possessive *his* than does the NTS-HGT at significant levels ($p < .0001$, log. likelihood value 7417.50). At this stage, it may be assumed that the majority of

these are context dependent, referring to the actual candidates in the presidential campaign in the USA in 2016 or individuals in the political discourse, as in (13)–(16).

(13) Watch Sean Hannity Expose Crooked Mueller and **His** Sleazy Dem Party Hit Squad

(14) BREAKING: Man Planning to KILL Trump was Hillary Clinton’s Close Friend! **His** Plan is SICK...

(15) Look How John McCain Responded When **His** Wife Joked About **His** Hair

(16) Comedian Dave Chapelle showed **his** support for Colin Kaepernick by wearing this shirt during his show in NYC

However, if we look solely at the singular third-person pronouns (*he/him/his/himself* vs. *she/her/hers/herself*), the datasets are surprisingly similar. Figure 6 visualizes the proportions of masculine and feminine pronouns, and shows that the feminine pronouns have a slightly higher proportion in the troll data (31.1%), while the respective figure for the NTS-HGT is 29.8%. The differences in the raw figures are statistically significant ($p < 0.0001$, log. likelihood value 154.42).

Figure 6. The proportions of masculine and feminine third-person pronouns in the datasets

This result is striking because masculine pronouns are more common than their feminine counterparts in all registers (Biber et al. 1999: 333). The most likely explanation for the slight preference in Figure 6 is caused by the context from which the troll data are derived, that is, the presidential campaign of 2016, which was a contest between a male and a female candidate. This is also reflected in the topics discussed in the troll data, as shown in (17)–(19).

(17) Hillary Clinton is trying to separate herself from Obama. I guess **she's** running from president.

(18) RT : Moms Demand Action volunteer drank 400 coffees for Hillary—and then **she** voted for her! ...

(19) Do not trust, Obama! The bandit Dilma is the worst kind **She** ended with Brazil **She** and Lula.

This section has offered further evidence of differences between the two datasets and has confirmed that the genuine personal messages (NTS-HGT) are closer to interactive conversational style than are the troll data. The former are pronoun dense and resemble conversation in which first- and second-person pronouns are more frequent, as observed by Biber et al. (1999: 333). Troll messages tend to resemble the news, where the actions and thoughts of people are frequently reported.

6. Discussion and conclusions

We have investigated sociolinguistic democratization by exploring anti-democratic discourse. We have focused on the linguistic properties of deception, misinformation, and false news as the antipodes of democratic ideals. All of these anti-democratic phenomena have been in existence for centuries, but it may be no exaggeration to claim that never has their role been more significant than today, when digital information can reach very large numbers of people or, alternatively, touch microscopic populations by means of skillfully designed and highly targeted digital social media campaigns. Hence, we need an expansion in empirical linguistic evidence of what actually characterizes the kind of malicious trolls that aim at influencing popular opinion and breeding discord in democratic societies. This information needs to be based on large sets of data, and to obtain this we need tested methodologies and tools. English corpus linguistics can substantially contribute to this topic methodologically, and it can provide much-needed socio-cultural contextualization for data mining endeavors. This study represents a first step towards this direction.

This article has attempted to demonstrate the usefulness of exploratory corpus analyses with robust statistics in the analysis of anti-democratic discourse. We have concentrated on high-frequency items and aimed at uncovering broad quantitative patterns in the datasets. The underlying assumption is based on sociolinguistics and on a notion that the use of frequent elements takes place below the level of linguistic awareness, which means that it is possible to uncover fundamental differences between two datasets. As of autumn 2019, the work towards describing the linguistic properties of online trolls and their (automated) detection has only started. This is evident in the number of publications posted in pre-publication archives. This work has used thematic information, sentiment analysis, keyword/n-gram information, and syntactic n-grams (e.g. Boyd et al. 2018; Im et al. 2019; Ghanem et al. 2019). While we have aimed at providing an overview of the linguistic properties of how troll data differ from genuine human accounts, we acknowledge that we simply do not yet have sufficient empirical evidence of the linguistic properties of troll data in English that originates from settings in which English is used as a non-native resource. By comparing troll data from non-native settings with baseline data that also originates from non-native contexts is an important aspect. It is a fact that a substantial share of English use online today is generated by non-native speakers. As an illustration, the share of English messages in the NTS corpus is over 30% of all the material outnumbering the main languages of the region (Laitinen et al. 2018). We have attempted to highlight the fact that since anti-democratic activities are unarguably global and often utilize English as the language for their global reach, studies of anti-democratic language use and attempts to detect trolls can benefit from evidence of different varieties of English. If the objective is to detect troll messages, it is not sufficient to distinguish non-native troll data from native language use, but to be able to distinguish non-native trolls from genuine outward non-native language use. Moreover, malicious trolling is not only a non-native English phenomenon, but it is safe to assume that trolling also takes place in native settings, and the knowledge generated from non-native settings could contribute to detecting malicious native trolls.

This study has used two datasets in which the common denominator is that the messages are in English, and as pointed out in the introduction, the only known parameter is that one of them consists of troll messages while the other contains genuine personal messages. The baseline data of NTS-HGT contain material from which social bots have been excluded, as specified in Section 3. However, we do not yet know if this material contains troll messages, because the authors do not have a functional troll detector. The troll data are publicly available data released by Twitter and contain material from known, state-backed information operations aimed at sowing discord in Western societies. The selection of these data is partly based on

convenience, since we have piloted the bot detection algorithm using the NTS data, and it enables us to increase data accuracy. It is clear that there are unknown parameters that could be controlled in future studies. These include for instance L1 differences or even time, since the NTS-HGT data are from 2017–2018 while the troll data are related to the 2016 elections in the U.S. Moreover, the news-headline style in the troll data suggests differences in text type, which could be investigated in more detail in future studies.

Our empirical observations show considerable quantitative differences between the datasets. The troll messages are shorter than the baseline material, and they also make use of a smaller number of lexical types and tokens, while the proportional share of the most frequent types is considerably lower than in the NTS-HGT material. Anti-democratic messages also resemble more formal registers in that they contain a higher proportion of nouns and elements used in noun phrases, such as the synthetic possessive 's, but this tendency is not visible in the use of articles, which distinguishes the quantitative profile of troll data. Personal messages are rich in verbs, pronouns, adverbs, and interjections, which all indicate an emphasis on action and outward interaction in particular.

The quantitative evidence also suggests that there seem to be consistent patterns constructing a propensity for self-references in the genuinely personal data, which is far less frequent in the anti-democratic data. In addition, the troll data contains slightly more negative utterances than does the NTS-HGT material.

The results suggest the feasibility of a range of future studies that might expand on the stock of linguistic features beyond personal pronouns to, for instance, modal and semi-modal auxiliaries or focus on the properties of noun and noun phrases in the troll data. Furthermore, it may well be useful to investigate aggregated grammatical items in terms of typological profiling or through multidimensional analysis. Another line to follow could well be a focus on content words, including keyword analysis and the expressions of sentiment in the datasets.

References:

- Alissandrakis, Aris, Nico Reski, Mikko Laitinen, Jukka Tyrkkö, Jonas Lundberg & Magnus Levin. 2019. Visualizing rich corpus data using virtual reality. In Hanna Parviainen, Mark Kaunisto & Päivi Pahta (eds.), *ICAME 39 Proceedings*. Helsinki: VARIENG e-series.
- Baker, Paul. 2009. The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics* 14:3, 312–337. doi 10.1075/ijcl.14.3.02bak.
- Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.
- Boyd, Ryan, Alexander Spangher, Adam Fourney, Besmira Nushi, Gireeja Ranade, James Pennebaker & Eric Horvitz. 2018. Characterizing the Internet Research Agency's social media operations during the 2016 U.S. presidential election using linguistic analyses. *PsyArXiv. October 1*. doi:10.31234/osf.io/ajh2q.
- Briscoe, Erica, D. Scott Appling & Heather Hayes. 2014. Cues to Deception in Social Media Communications. *47th Hawaii International Conference on System Science*, 1435–1442. IEEE. doi10.1109/HICSS.2014.186.
- Burgoon, Judee, J.P. Blair, Tiantian Qin & Jay F. Nunamaker, Jr. 2003. Detecting deception through linguistic analysis. In Hsinchun Chen, Richard Miranda, Daniel D. Zeng, Chris Demchak, Jenny Schroeder & Therani Madhusudan (eds.), *Intelligence and Security Informatics. ISI 2003. Lecture Notes in Computer Science*, 91–101. Vol 2665. Berlin, Heidelberg: Springer.
- CNN 2019. Mueller calls out high profile Americans who unwittingly helped Russian trolls. Donnie O'Sullivan, April 19, 2019. Available at <https://edition.cnn.com/2019/04/18/tech/internet-research-agency-mueller-report/index.html>. Accessed 15 May, 2019.
- Coats, Steven. 2017. European language ecology and bilingualism with English on Twitter. – Ciara Wigham & Egon Stemle (eds.), *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities*. Bozen/Bolzano: Eurac Research, 35–38.
- Cogo, Alesia & Martin Dewey. 2006. Efficiency in ELF communication. *Nordic Journal of English Studies* 5:2, 59–93.
- Coles, Bryn Alexander & Melanie West. 2016. Trolling the trolls: Online forum users constructions of the nature and properties of trolling. *Computers in Human Behavior* 60, 233–244. doi <https://doi.org/10.1016/j.chb.2016.02.070>.
- Farrelly, Michael & Elena Seoane. 2012. Democratization. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), *The Oxford Handbook of the History of English*, 392–401. Oxford: Oxford University Press.
- Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer & Alessandro Flammini. 2014. The rise of social bots. *Communications of the ACM*, 59:7, 96–104. doi10.1145/2818717.
- Ghanem, Bilal, Davide Buscaldi & Paolo Rosso. 2019. TexTrolls: Identifying Russian Trolls on Twitter from a Textual Perspective. arXiv:1910.01340v1 [cs.CL].
- Guardian 2018. Tom Watson: UK on frontier of 'new cold war' that Russia is winning. The Guardian, 23 Sept. 2018. Available at <https://www.theguardian.com/politics/2018/sep/23/tom-watson-uk-on-frontier-of-new-cold-war-that-russia-is-winning>. Accessed 9 Oct 2018.
- Gupta, Aditi, Hemank Lamba & Ponnurangam Kumaraguru. 2013. \$1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on Twitter. *eCrime Researchers Summit. IEEE*, 1–12. doi 10.1109/eCRS.2013.6805772.
- Honey, Courtenay & Susan Herring. 2009. Beyond microblogging: conversation and collaboration via Twitter. In *the 42nd Hawai'i International Conference on System Sciences (HICSS-42)*. Los Alamitos, California: IEEE Press. doi 10.1109/HICSS.2009.89.
- Huang, Yuan, Diansheng Guo, Alice Kasakoff & Jack Grieve. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59, 244–255. doi 10.1016/j.compenvurbsys.2015.12.003.

- Im, Jane, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens & Eric Gilbert. 2019. Still out there: Modeling and Identifying Russian Troll Accounts on Twitter. arXiv:1901.11162 [cs.SI].
- Knight, Dawn, Svenja Adolphs & Ronald Carter. 2014. CANELC: constructing an e-language corpus. *Corpora* 9:1, 29–56. doi 10.3366/cor.2014.0050.
- Labov, William. 2007. Transmission and diffusion. *Language* 83:2, 344–387.
- Laitinen, Mikko & Jonas Lundberg. In press. ELF, language change and social networks: Evidence from real-time social media data. In Anna Mauranen & Svetlana Vetchinnikova (eds.), *Language Change: The Impact of English as a Lingua Franca*. Cambridge: Cambridge University Press.
- Laitinen, Mikko, Jonas Lundberg, Magnus Levin & Rafael Martins. 2018. The Nordic Tweet Stream: A Dynamic Real-Time Monitor Corpus of Big and Rich Language Data, *Proc. of Digital Humanities in the Nordic Countries 3rd Conference*, Helsinki, Finland, March 7-9, 2018, CEUR-WS.org, online CEUR-WS.org/Vol-2084/short10.pdf.
- Laitinen, Mikko, Jonas Lundberg, Magnus Levin & Alexander Lakaw. 2017. Revisiting weak ties: using present-day social media data in variationist studies. In Tanja Säily, Minna Palander-Collin, Arja Nurmi, & Anita Auer (eds.), *Exploring Future Paths for Historical Sociolinguistics*, 303–325. Amsterdam: John Benjamins. doi 10.1075/ahs.7.12lai.
- Laitinen, Mikko. 2008. Sociolinguistic patterns in grammaticalisation: HE, THEY, and THOSE in human indefinite reference. *Language Variation and Change* 20:1, 155–185. doi <https://doi.org/10.1017/S0954394508000045>.
- Lamanna Fabio, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves & José J. Ramasco. 2018. Immigrant community integration in world cities. *PLoS ONE* 13(3): e0191612. <https://doi.org/10.1371/journal.pone.0191612>
- Leech, Geoffrey, Marianne Hundt, Christian Mair & Nicholas Smith. 2009. *Change in Contemporary English. A Grammatical Study*. Cambridge: Cambridge University Press.
- Lundberg, Jonas, Jonas Nordqvist & Mikko Laitinen. 2019. Towards a language independent Twitter bot detector. In Costanza Navarretta, Manex Agirrezabal & Bente Maegaard (eds.), *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference Copenhagen, Denmark, March 5–8, 2019*, 308–319. Available at <http://ceur-ws.org/Vol-2364/>.
- Mair, Christian. 1997. Parallel corpora: A real time approach to the study of language change in progress. In Magnus Ljung (ed.), *Corpus-Based Studies in English*, 195–209. Amsterdam: Rodopi.
- Mauranen, Anna. 2018. Conceptualising ELF. In Jennifer Jenkins, Will Baker & Martin Dewey (eds.), *The Routledge Handbook of English as a Lingua Franca*. London: Routledge.
- Mauranen, Anna. 2012. *Exploring ELF: Academic English Shaped by Non-Native Speakers*. Cambridge: Cambridge University Press.
- Morstatter, Fred, Liang Wu, Tahora H. Nazer, Kathleen M. Carley & Huan Liu. 2016. A new approach to bot detection: striking the balance between precision and recall. *International Conference on Advances in Social Networks Analysis and Mining*, 533–540. doi 10.1109/ASONAM.2016.7752287.
- Myhill, John. 1995. Change and continuity in the functions of the American English Modals. *Linguistics* 33, 157–211.
- Oakes, Michael. 2009. Corpus linguistics and language variation. In Paul Baker (ed.), *Contemporary Approaches to Corpus Linguistics*, 159–183. London: Continuum.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

- Romaine, Suzanne. 2001. A corpus-based view of gender in British and American English. In Marlis Hellinger & Hadumod Bussmann (eds.), *Gender across Languages: The Linguistic Representation of Women and Men*, 153–174. Vol. 1. Amsterdam: John Benjamins.
- Tagliamonte, Sali A. 2012. *Variationist Sociolinguistics: Change, Observation, Interpretation*. Oxford: Wiley-Blackwell.