palgrave communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

https://doi.org/10.1057/s41599-019-0280-3 OPEN

How character limit affects language usage in tweets

Arnout B. Boot[1], Erik Tjong Kim Sang[2], Katinka Dijkstra[1] & Rolf A. Zwaan[1]

ABSTRACT In November 2017 Twitter doubled the available character space from 140 to 280 characters. This provided an opportunity for researchers to investigate the linguistic effects of length constraints in online communication. We asked whether the character limit change (CLC) affected language usage in Dutch tweets and hypothesized that there would be a reduction in the need for character-conserving writing styles. Pre-CLC tweets were compared with post-CLC tweets. Three separate analyses were performed: (I) general analysis: the number of characters, words, and sentences per tweet, as well as the average word and sentence length. (II) Token analysis: the relative frequency of tokens and bigrams; (III) part-of-speech analysis: the grammatical structure of the sentences in tweets (i.e., adjectives, adverbs, articles, conjunctives, interjections, nouns, prepositions, pronouns, and verbs); pre-CLC tweets showed relatively more textisms, which are used to abbreviate and conserve character space. Consequently, they represent more informal language usage (e.g., internet slang); in turn, post-CLC tweets contained relatively more articles, conjunctions, and prepositions. The results show that online language producers adapt their texts to overcome limit constraints.

[1] Erasmus University Rotterdam, Mandeville building, room T16-03, Burgemeester Oudlaan 50, Rotterdam, NL 3062 PA, The Netherlands. [2] Netherlands eScience Center, Amsterdam, The Netherlands. Correspondence and requests for materials should be addressed to A.B.B. (email: boot@essb.eur.nl)

Introduction

Spontaneous linguistic communication is typically unrestrained in terms of the length of utterances but in some situations there are constraints on utterance length. For example, there are word count limitations to newspaper headlines, advertisements, journalistic articles, student papers, and scholarly manuscripts. These limitations are sometimes so restrictive that they impact sentence structure and content and word forms. For instance, the advent of the telegraph, in which words were literally at a premium, necessitated an elliptic style that has become known as telegram style of *telegraphese*, which is viewed as a normal expressive form of language (Barton, 1998; Isserlin, 1985; Tesak and Dittmann, 2009). A more contemporary example of an elliptic style is *textese*, which is often used in modern text messages (Drouin and Driver, 2014).

Textese and telegraphese are both characterized by an imposed limit constraint (Barton, 1998; Drouin and Driver, 2014; Isserlin, 1985; Tesak and Dittmann, 2009). However, a crucial difference is the nature of the length restriction: In telegrams, the costs are related to the number of words and not the number of characters. In other words, a cost-effective telegram contains as few words as possible. In text messages, on the other hand, one is obliged to conserve character space, which results in a different practice of economy (Frehner, 2008). Character reduction as performed in textese, can be achieved not only by minimizing the number of words but also by abbreviating words and using shorter synonyms and symbols. Textese has been called 'squeeze text', which well reflects its grammatical features (Carrington, 2004).

The character-reducing strategies inherent to textese are referred to as *textisms* (Carrington, 2004; Lyddy et al. 2014). They evolved not only to save character space but also to reduce typing efforts. Textisms reduce character use without compromising the conveyed meaning and even add meaning in some cases. This includes acronyms (e.g., *LOL* for 'laugh out loud'), emoticons (e.g., ☺ instead of 'I am happy'), accent stylizations (e.g., slang terms such as *gonna*), nonconventional spellings (e.g., *gudnite*), homophones (e.g., *gr8* and *c u*), shortenings (e.g., *pic* as in 'picture'), contractions (e.g., *thx* for 'thanks'), and omission of punctuation (Carrington, 2004; De Jonge and Kemp, 2012; Ling and Baron, 2007; Plester et al., 2009; Tagliamonte and Denis, 2008; Thurlow and Brown, 2003; Varnhagen et al., 2010).

Another strategy to reduce character usage is the omission of certain part-of-speech (POS) categories. The basic elements of a sentence are subject, verb, and object (SVO or SOV; Koster, 1975). The SVO structure, comprises (pro)nouns and a verb. For example, '*Tom ate lunch*'. The main components of the SVO structure are unlikely to be omitted. In contrast, the POS categories that modify the basic structure and introduce additional information are more likely to be excluded. In textese and telegraphese, articles and conjunctions are often excluded (Carrington, 2004; Oosterhof and Rawoens, 2017). Consistent with this intuition, eyetracking studies of reading have shown that function words such as articles and prepositions are often skipped in normal reading because these words are both short and highly predictable from context (Rayner et al., 2011). A reader can even fill in omitted articles and conjunctions. For example, '*car broke down stopped in middle of road*'. Although the overall readability is compromised, the message is still clear. Therefore, if words have to be omitted to reduce character usage, they are likely to be function words. However, other words can also be omitted, leaving out information. For example, '*the car broke down*' instead of '*the car broke down and stopped in the middle of the road*'. In this case, additional information is being withheld. Generally, this means limit constraints might also affect sentence structure.

An example of a contemporary platform that might necessitate elliptic writing strategies is Twitter, an online microblogging platform which imposes a message-length limit to its users. On November 8th 2017, Twitter doubled the character limit from 140 characters to 280 characters[1]; we will refer to this as the character limit change (CLC). After a trial period in September, Twitter observed that 9% of English tweets hit the previous limit of 140 characters, whereas only 1% of tweets reached the new 280-character limit (Rosen, 2017). Doubling the character limit was thought to prevent a group of users from 'cramming their thoughts' (Rosen and Ihara, 2017). Furthermore, only 2% of trial tweets surpassed 190 characters, indicating that many users used merely a few more characters than had previously been possible. When Twitter announced the upcoming CLC the community responded ambivalently. Some users appreciated the increased tweet length, having more space to express their thoughts, whereas others claimed it would harm the tweets' brevity and to-the-point characteristics (Watson, 2017).

The doubling of the maximum tweet length provides for an interesting opportunity to investigate the effects of a relaxation of length constraints on linguistic messaging. What happened to the average length of tweets? And more interestingly, how did CLC impact the structure and word usage in tweets?

The need for an economy of expression decreased post-CLC. Therefore, our first hypothesis states that post-CLC tweets contain relatively less textisms, such as abbreviations, contractions, symbols, or other 'space-savers'. In addition, we hypothesize that the CLC affected the POS structure of the tweets, containing relatively more adjectives, adverbs, articles, conjunctions, and prepositions. These POS categories carry additional information about the situation being described, the referential situation; such as features of entities, the temporal order of events, locations of events or objects, and causal connections between events (Zwaan and Radvansky, 1998). This structural change also entails that sentences will be longer, with more words per sentence.

Gligorić et al. (2018) compared pre and post-CLC tweets with a length of approximately 140 characters. They found that pre-CLC tweets in this character range comprise relatively more abbreviations and contractions, and fewer definite articles. In the current study, we used a different approach that adds complementary value to the previous findings: we performed a content analysis on a dataset of approximately 1.5 million Dutch tweets including all ranges (i.e., 1–140 and 1–280), instead of selecting tweets within a specific character range. The dataset comprises Dutch tweets that were created between 25 October 2017 and 21 November 2017, in other words two weeks prior to and two weeks after the CLC.

We performed a general analysis to investigate changes in the number of characters, words, sentences, emojis, punctuation marks, digits, and URLs. To test the first hypothesis, we performed token and bigram analyses to detect all changes in the relative frequencies of tokens (i.e., individual words, punctuation marks, numbers, special characters, and symbols) and bigrams (i.e., two-word sequences). These changes in relative frequencies could then be utilized to extract the tokens that were especially affected by the CLC. In addition, a POS analysis was performed to test the second hypothesis; that is, whether the CLC affected the POS structure of the sentences. An example of each investigated POS category is presented in Table 1.

Method

Apparatus. The data collection, pre-processing, quantitative analysis, figures, token analysis, bigram analysis, and POS analysis

Table 1 Part-of-speech (POS) categories of interest

POS category	Example	Function
Adjective	cold, happy, young, two, fun	Describes, modifies or gives more information about a noun or pronoun
Adverb	slowly, very, always, well, too	Modifies a verb, an adjective or another adverb. It tells 'how' (often) and 'when'
Article	it, a, an[a]	Defines a noun as definite or indefinite
Conjunctive	and, or, but, because, yet, so	Joins two words, ideas, phrases together and shows how they are connected
Interjection	haha, wow, hey, yes, oh	Expression of a strong emotion with a brief exclamation
Noun	house, chair, dog, Mary, Tom	Name of a person, place, or any object
Preposition	at, on, in, from, with, about	Shows the relationship of a noun or pronoun to another word
Pronoun	I, you, it, we, them, those	Reference to a person or object
Verb	are, is, go, speak, live, eat	Depicts an action or state of being

These word classes can be applied to the Dutch language as well
[a]The Dutch definite article words also distinguish masculine/feminine nouns (i.e., 'de') and neuter nouns ('het')

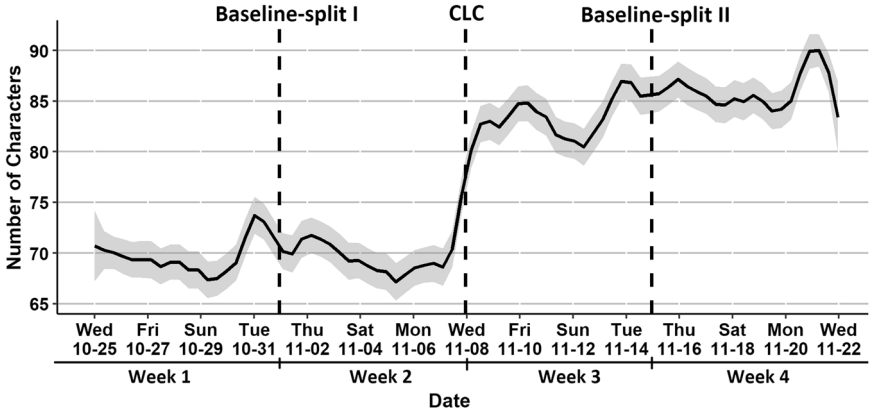

Fig. 1 Moving average and standard error of the character usage over time, which shows an increase in character usage post-CLC and an additional increase between week 3 and 4. Each tick marks the absolute beginning of the day (i.e., 12:00 a.m.). The time frames indicate the comparative analyses: week 1 with week 2 (Baseline-split I), week 3 with week 4 (Baseline-split II), and week 1 and 2 with week 3 and 4 (CLC)

were performed using Rstudio (RStudio Team, 2016). The R packages that were used are: 'BSDA', 'dplyr', 'ggplot', 'grid', 'kableExtra', 'knitr', 'lubridate', 'NLP', 'openNLP', 'quanteda', 'R-basic', 'rtweet', 'stringr', 'tidytext', 'tm' (Arnholt and Evans, 2017; Benoit, 2018; Feinerer and Hornik, 2017; Grolemund and Wickham, 2011; Hornik, 2016; Hornik, 2017; Kearney, 2017; R Core Team, 2018; Silge and Robinson, 2016; Wickham, 2016; Wickham, 2017; Xie, 2018; Zhu, 2018).

Period of interest. The CLC occurred on 8 November 2017 at 00:00 a.m. (UTC). The dataset comprises Dutch tweets that were created within two weeks pre-CLC and two weeks post-CLC (i.e., from 10-25-2017 to 11-21-2017). This period is subdivided into *week 1, week 2, week 3*, and *week 4* (see Fig. 1). To analyze the effect of the CLC we compared the language usage in 'week 1 and week 2' with the language usage in 'week 3 and week 4'. To distinguish the CLC effect from natural-event effects, a control comparison was devised: the difference in language usage between week 1 and week 2, referred to as *Baseline-split I*. Furthermore, the CLC could have initiated a trend in the language usage that evolved as more users became familiar with the new limit. This trend could be shown by comparing week 3 with week 4, referred to as *Baseline-split II*.

Data collection. The website[2] *twiqs.nl* was used as a means to collect tweet-ids[3], this website provides researchers with metadata from a (third-party-collected) corpus of Dutch tweets (Tjong Kim Sang and Van den Bosch, 2013). The tweet-ids allow for the collection of tweets from the Twitter API that are older than 9 days (i.e., the historical limit when requesting tweets based on a search query). The R-package 'rtweet' and complementary 'lookup_status' function were used to collect tweets in JSON format. The JSON file comprises a table with the tweets' information, such as the creation date, the tweet text, and the source (i.e., type of Twitter client).

Data cleaning and preprocessing. The JSON[4] files were converted into an R data frame object. Non-Dutch tweets, retweets, and automated tweets (e.g., forecast-, advertisement-relatea, and traffic-related tweets) were removed. In addition, we excluded tweets based on three user-related criteria: (1) we removed tweets that belonged to the top 0.5 percentile of user activity because we considered them non-representative of the normal user population, such as users who created more than 2000 tweets within four weeks. (2) Tweets from users with early access to the 280 limit were removed. (3) Tweets from users who were not represented in both pre and post-CLC datasets were removed, this procedure ensured a consistent user sample over time (within-group design, $N_{users} = 109{,}661$). All cleaning procedures and corresponding exclusion numbers are presented in Table 2.

The tweet texts were converted to ASCII encoding. URLs, line breaks, tweet headers, screen names, and references to screen names were removed. URLs add to the character count when located within the tweet. However, URLs do not add to the character count when they are located at the end of a tweet. To prevent a misrepresentation of the actual character limit that users had to deal with, tweets with URLs (but not media URLs such as added pictures or videos) were excluded.

Table 2 Dataset exclusions and inclusions

	Pre-CLC		Post-CLC	
Tweet type excluded	**Number of tweets**	**Proportion**	**Number of tweets**	**Proportion**
Non-representative Twitter clients (e.g., bots/automated tweets)	1329457	0.27	1361462	0.26
Upper 0.5% of user activity	928627	0.19	1095653	0.21
URL/ad related	573654	0.12	581951	0.11
Non-Dutch	389315	0.08	421593	0.08
Retweets	660791	0.14	740885	0.14
Tweets without words (<1%)	725	0	735	0
Users absent in pre or post-CLC dataset[a]	250464	0.05	235035	0.05
Included tweets (*N*)	744673	0.15	764642	0.15

[a]Users not represented in both pre and post-CLC datasets (or users with early access to the 280 limit)

Table 3 Tweet features pre and post-CLC

	Pre-CLC			Post-CLC			Difference	
Number of:[a]	**Mean**	**SD**	**99% CI**	**Mean**	**SD**	**99% CI**	**Absolute**	**Relative (%)**
Tweets per user	6.79	10.60	[6.71, 6.87]	6.97	10.63	[6.89, 7.06]	+0.18	+2.65
Characters	70.08	39.06	[69.96, 70.19]	84.82	60.63	[84.64, 84.99]	+14.74	+21.03
Words	11.89	6.64	[11.87, 11.91]	14.21	10.01	[14.18, 14.24]	+2.32	+19.51
Sentences	1.55	0.80	[1.55, 1.56]	1.70	1.02	[1.70, 1.70]	+0.15	+9.68
Characters per word	4.77	1.27	[4.77, 4.77]	4.81	1.48	[4.80, 4.81]	+0.04	+0.84
Characters per sentence	49.51	30.32	[49.42, 49.6]	53.42	36.32	[53.31, 53.53]	+3.91	+7.90
Words per sentence	8.49	5.27	[8.47, 8.50]	9.07	6.17	[9.05, 9.09]	+0.58	+6.83
Emojis	0.27	0.93	[0.27, 0.27]	0.28	1.12	[0.28, 0.28]	+0.01	+3.70
Digit characters	0.45	1.41	[0.45, 0.45]	0.53	1.84	[0.53, 0.54]	+0.08	+17.78
Numbers	0.27	0.77	[0.27, 0.27]	0.29	0.90	[0.29, 0.30]	+0.02	+7.41
Punctuation marks	2.41	2.30	[2.40, 2.42]	2.85	3.21	[2.84, 2.86]	+0.44	+18.26
URLs	0.11	0.31	[0.11, 0.11]	0.13	0.34	[0.13, 0.13]	+0.02	+18.18

99% CIs were implemented, as opposed to the traditional 95% CI, to reduce the chance of type I errors. The CIs are narrow because the sample size is very large
[a]All feature means were computed per tweet, except for the number of tweets, which was computed per user

Token and bigram analysis. The R package[5] 'quanteda' was used to tokenize the tweet texts into tokens (i.e., isolated words, punctuation marks, and numbers) and bigrams. In addition, token-frequency-matrices were computed with: the frequency pre-CLC [*f*(token pre)], the relative frequency pre-CLC[*P* (token pre)], the frequency post-CLC [*f*(token post)], the relative frequency post-CLC and *T*-scores. The *T*-test is similar to a standard *T*-statistic and computes the statistical difference between means (i.e., the relative word frequencies). Negative *T*-scores indicate a relatively higher occurrence of a token pre-CLC, whereas positive *T*-scores indicate a relatively higher occurrence of a token post-CLC. The *T*-score equation used in the analysis is presented as Eq. (1) and (2). *N* is the total number of tokens per dataset (i.e., pre and post-CLC). This equation is based on the method for linguistic computations by Church et al. (1991; Tjong Kim Sang, 2011).

$$T = \frac{P(\text{token post}) - P(\text{token pre})}{\sqrt{\sigma^2(P(\text{token post})) + \sigma^2(P(\text{token pre}))}} \quad (1)$$

$$\approx \frac{\frac{f(\text{token post})}{N_{\text{pre}}} - \frac{f(\text{token pre})}{N_{\text{post}}}}{\sqrt{\frac{f(\text{token post})}{N^2_{\text{post}}} + \frac{f(\text{token pre})}{N^2_{\text{pre}}}}} \quad (2)$$

Part-of-speech (POS) analysis. The R package[6] 'openNLP' was used to classify and count POS categories in the tweets (i.e., adjectives, adverbs, articles, conjunctives, interjections, nouns, numeral, prepositions, pronouns, punctuation, verbs, and miscellaneous). The POS tagger operates using a maximum entropy (maxent) probability model in order to predict the POS category based on contextual features (Ratnaparkhi, 1996). The Dutch maxent model used for the POS classification was trained on CoNLL-X Alpino Dutch Treebank data (Buchholz and Marsi, 2006; Van der Beek et al., 2002). The openNLP POS model has been reported with an accuracy rating of 87.3% when used for English social media data (Horsmann et al., 2015). An ostensible limitation of the current study is the reliability of the POS tagger. However, similar analyses were performed for both pre-CLC and post-CLC datasets, meaning the accuracy of the POS tagger should be consistent over both datasets. Therefore, we assume there are no systematic confounds.

Statistical interpretation. The large sample size ($N = 1{,}516{,}425$) is an approximation of the population size; this means that the standard errors are low and the confidence intervals (CI) are narrow. 99% CIs were implemented, as opposed to the commonly used 95% CI, to reduce the chance of type I errors.

Results

The results comprise three components: (1) General statistics–the CLC induced differences across multiple tweet features, (2) token (i.e., unigram) and bigram analyses to test the first hypothesis, and (3) POS analysis to test the second hypothesis.

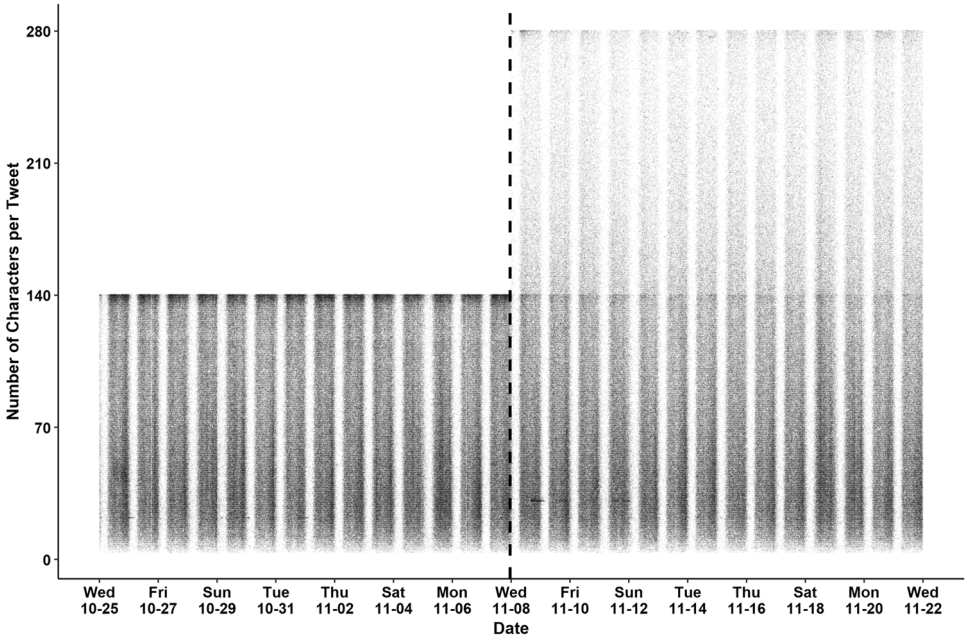

Fig. 2 Character usage over time. This scatterplot displays the number of characters in each tweet ($n = 1{,}509{,}315$) over time. The reference line indicates the CLC. The observations show an increase in character usage post-CLC, fewer tweets accumulating near 140 characters post-CLC, the day/night cycle of tweet behavior, and a small proportion of tweets that were still limited by the 140-limit post-CLC (outdated Twitter client versions)

General statistics. After the CLC, the average tweet length increased. Table 3 contains descriptive information about different tweet features such as character and word count. This table also provides the absolute and relative differences between pre and post-CLC tweets. All tweet features increased in frequency. Furthermore, the standard deviations of all length features increased, indicating an increase in variability. This suggests some users took advantage of the additional character space, whereas others continued to use fewer than 140 characters.

Figure 1 shows that the average character usage increased immediately after the CLC. In addition, the character usage also increased from week 3 to week 4, suggesting that some users became familiar with the 280-limit in the week after the CLC. Figure 2 provides an overview of all observations and shows an increase in character usage from pre to post-CLC time frames. This figure also shows the day/night cycle in Twitter activity, a small proportion of users who were still limited to 140 characters after the CLC (due to outdated Twitter client versions), an initial increase in the amount of tweets near the 280-limit, and a decrease in the amount of tweets near the 280-limit as compared to the 140-limit. Figure 3 displays the character (3a), word (3b), and sentence (3c) usage over time, which show a similar increase in tweet length. Figure 4a displays the number of characters per word (i.e., word length) over time. The average word length remained unaffected by the CLC, except for a temporary increase the first day after the CLC. Figure 4b, c present an increase in sentence length after the CLC, this suggests a syntactic change in sentence structure.

Figure 5 shows a large amount of pre-CLC tweets (15.48%) within the upper range of 121–140 characters. In comparison, a much smaller proportion of post-CLC tweets (1.73%) are within the upper range of 261–280 characters. Alternatively, the percentage of pre-CLC tweets near the pre-CLC limit (i.e., 138–140 characters) is 4.73%, whereas the post-CLC limit (i.e., 278–280 characters) comprises just 0.48% of post-CLC tweets. In other words, doubling the character limit appears to have decreased the hindrance by a factor of ten.

Figure 6 shows the distribution of word usage in tweets pre and post-CLC. Again, it is shown that with the 140-characters limit, a group of users were constrained. This group was forced to use about 15 to 25 words, indicated by the relative increase of pre-CLC tweets around 20 words. Interestingly, the distribution of the number of words in post-CLC tweets is more right skewed and displays a gradually decreasing distribution. In contrast, the post-CLC character usage in Fig. 5 shows small increase at the 280-characters limit.

Token and bigram analyses. To test our first hypothesis, which states that the CLC reduced the use of textisms or other character-saving strategies in tweets, we performed token and bigram analyses. Firstly, the tweet texts were separated into tokens (i.e., words, symbols, numbers and punctuation marks). For each token the relative frequency pre-CLC was compared to the relative frequency post-CLC, thus revealing any effects of the CLC on the use of any token. This comparison of pre and post-CLC percentage was revealed in the form of a *T*-score, see Eqs. (1) and (2) in the method section. Negative *T*-scores indicate a relatively higher frequency pre-CLC, whereas positive T-scores indicate a relatively higher frequency post-CLC. The total number of tokens in the pre-CLC tweets is 10,596,787 including 321,165 unique tokens. The total number of tokens in the post-CLC tweets is 12,976,118 which comprises 367,896 unique tokens. For each unique token three *T*-scores were computed, which indicates to what extent the relative frequency was affected by Baseline-split I, Baseline-split II and the CLC, respectively (see Fig. 1).

Figure 7 presents the distribution of the *T*-scores after removal of low frequency tokens, which shows the CLC had an independent effect on the language usage as compared to the baseline variance. Particularly, the CLC effect induced more *T*-scores <-4 and >4, as indicated by the reference lines. In addition, the *T*-score distribution of the Baseline-split II comparison shows an intermediate position between Baseline-split I and the CLC. That is, more variance in token usage as compared to Baseline-split I, but less variance in token usage as

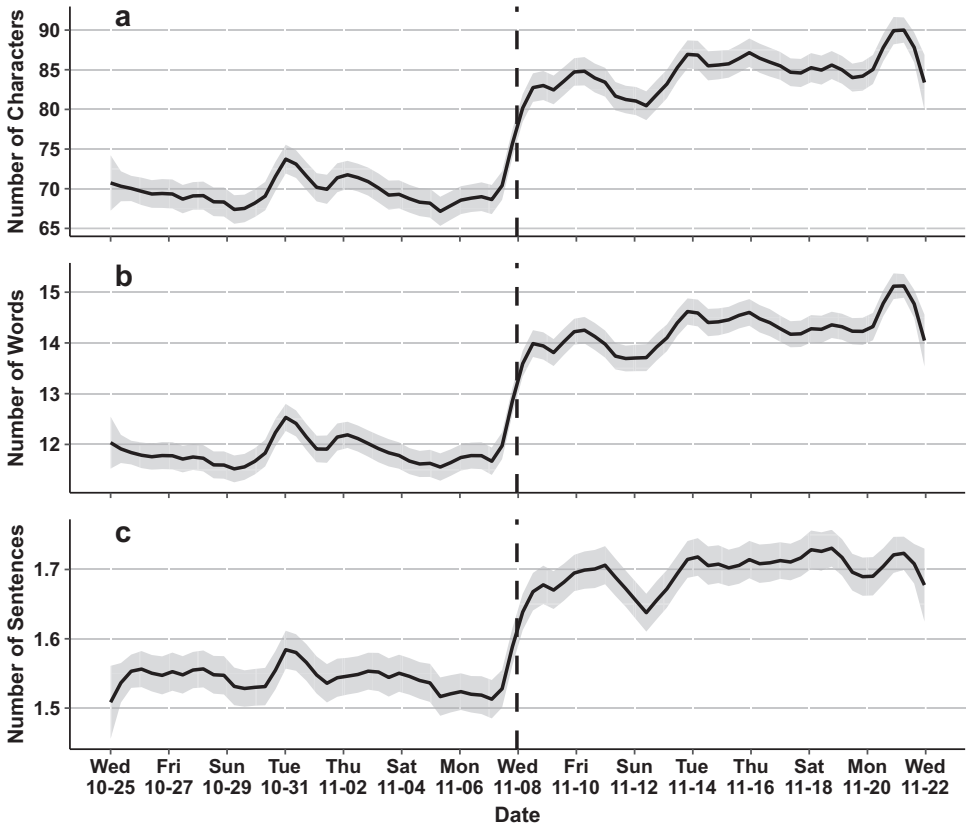

Fig. 3 Moving averages for the number of characters **a**, words **b**, and sentences **c**, including standard errors. The reference line indicates the CLC. Each tick marks the absolute beginning of the day (i.e., 12:00 a.m.). The moving averages show an increase in tweet length post-CLC. Character, word, and sentence usage display a similar increase post-CLC

compared to the CLC. Therefore, Baseline-split II (i.e., comparison between week 3 and week 4) could suggests a subsequent trend of the CLC. In other words, a gradual change in the language usage as more users became familiar with the new limit.

To minimize natural-event-related confounds the *T*-score range, indicated by the reference lines in Fig. 7, was utilized as a cutoff rule. That is, tokens within the range of −4 to 4 were excluded, because this range of *T*-scores can be ascribed to baseline variance, as opposed to CLC-dependent variance. Furthermore, we removed tokens that showed greater variance for Baseline-split I as compared to the CLC. A similar procedure was performed with bigrams, resulting in a *T*-score cutoff-rule of −2 to 2, see Fig. 8. Tables 4–7 present a subset of tokens and bigrams of which occurrences were the most affected by the CLC. Each individual token or bigram in these tables are accompanied by three related *T*-scores: Baseline-split I, Baseline-split II, and CLC. These *T*-scores can be used to compare the CLC effect with Baseline-split I and Baseline-split II, for each individual token or bigram.

The tokens that occurred relatively less frequently post-CLC are presented in Table 4. These tokens comprise: symbols (e.g., *&*, >, */*, +, ^, =), numerals (e.g., *1, 2, 3*) acronyms, shortenings and contractions (e.g., *t, k, ff, ni, mn, nie, jy, gwn, s, lol*; which refer to: *het, dat, ok/ik, even, niet, hem, niet, jij, gewoon, is, laugh out loud*; translations: *it, that, ok/I, for a bit*[7], *not, my, not, you, just, is*), punctuation marks (e.g., *! ?* : ; but not the period and comma), pronouns (e.g., *ik, jij, hem, hij, me, je, jou*; translations: *I, you, him, he, me, you/your*), opinion-related adjectives/adverbs (e.g., echt, *lekker, mooi, goed, nieuwe, niks, leuk, zeker, mooie, super*; translations: *really, nice/tasty, nice/beautiful, good, new, nothing, nice/beautiful, nice/nicely, sure, nice, super*), and interjection words (e.g., *ja, haha, nee, man, hoor, nou, hahaha, he, jaa, wow, jaaa, ok, fuck, shit, wtf*; translations: *yes, haha, no, man, you know, well, hahaha, hey/huh, yeah, wow, okay*). In summary, the words that occurred relatively more frequently pre-CLC represent mainly informal language use, such as contractions, unconventional spellings, symbols and profanity.

Table 5 presents tokens that occurred relatively more frequently *post*-CLC, these tokens comprise: articles (i.e., *de, het, een*; translations: feminine/masculine *the*, neuter *the*, *a(n)*), conjunctions (e.g., *en, of, omdat, want, zodat*; translations: *and, or, because, because, so that*), prepositions (e.g., *door, in, om, met, over, tijdens, aan, tot*; translations: *through/by, in, for/at, with, about/over, during, to/on, until*), auxiliary and linking verbs (e.g., worden, *hebben, zijn, moeten, kunnen, maken, willen*; translations: *become, have, are, must, can, make, want*). Overall, the tokens that occurred relatively more frequently post-CLC represent more formal language usage as compared to the pre-CLC tokens in Table 4.

Table 6 presents bigrams that occurred relatively more frequently pre-CLC. These bigrams mainly comprise *personal pronoun + verb* combinations (i.e., *ik ga, ik heb, ik ben, ik wil, ik dacht, heb je, ik moet, denk ik, ik kan, ik kom, ik had, ik was*; translations: *I am going, I have, I am, I want, I thought, have you, I must, I think, I can, I come, I had, I was*). Again, the results suggest that there was relatively more informal language usage, that is, relatively more frequent occurrences of self-referential language, which implies a more personal and subjective language usage.

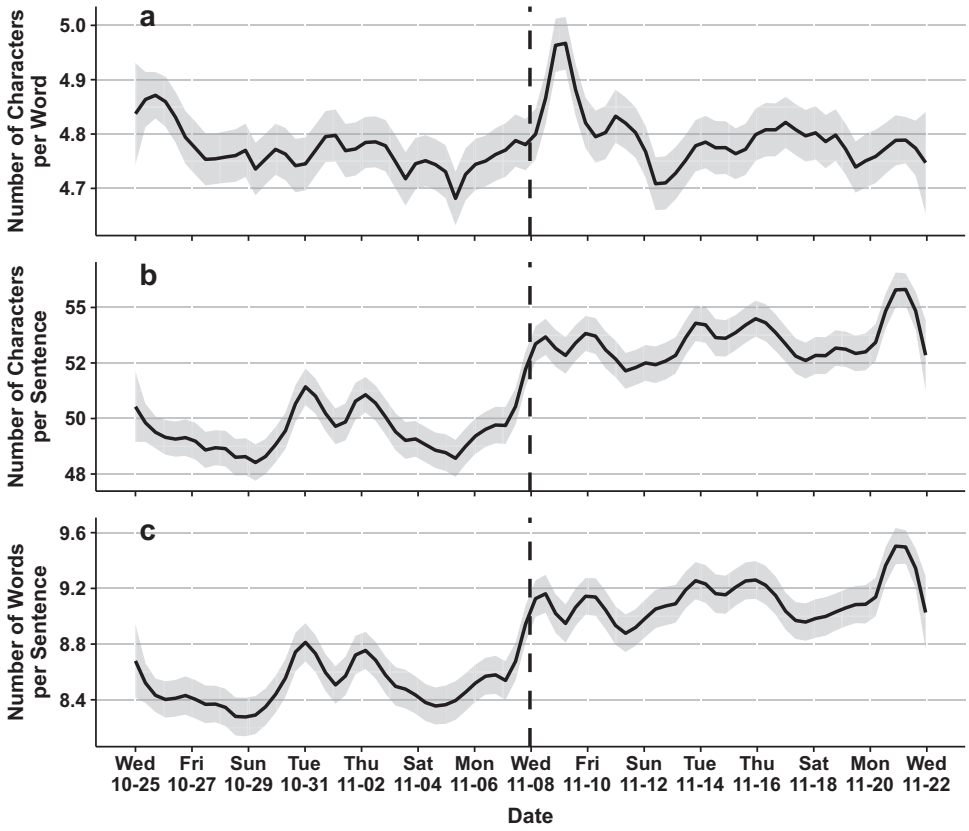

Fig. 4 Moving averages for the number of characters per word **a**, characters per sentence **b**, and words per sentence **c**, including standard errors. The reference line indicates the CLC. Each tick marks the absolute beginning of the day (i.e., 12:00 a.m.). Word length increased temporarily post-CLC but then decreased to the previous level. Sentences contained more characters and words post-CLC

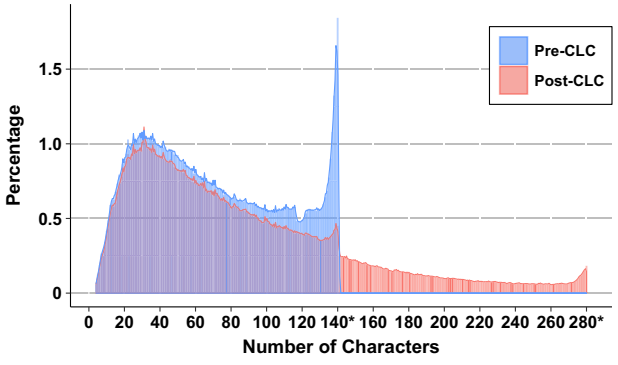

Fig. 5 Character-usage distribution; pre and post-CLC. This density distribution shows a large proportion of pre-CLC tweets within the upper range of 120–140 characters, whereas the proportion of post-CLC tweets within the upper range of 260–280 characters was reduced by a factor of ten

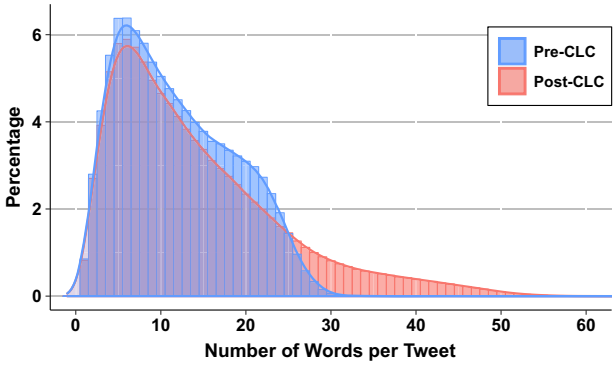

Fig. 6 Word-usage distribution; pre and post-CLC. This density distribution shows that in pre-CLC tweets there were relatively more tweets within the range of 15–25 words, whereas post-CLC tweets shows a gradually decreasing distribution and double the maximum word usage

The bigrams that occurred relatively more frequently post-CLC, in Table 7, comprise mainly prepositional phrases or preposition + article combinations (e.g., *van de, van het, door de, naar het, van een, om de, over de, aan de, over het, in het, met het, met de, om het, bij het, om een, voor het*; translations: *from the, from the, by the, to the, from a, about the, over the, in the, with the, about/over the, by the, around the, for the*), suggesting more detailed descriptions of the situation that is referred to in the tweets. Importantly, the introduction of extra prepositions can also explain the increase in sentence length after the CLC.

POS analysis. The second hypothesis about a potential increase in the use of adjectives, adverbs, articles, conjunctions, and prepositions, was tested using a POS analysis. Table 8 displays the relative frequencies of POS categories. Figure 9 presents the relative differences in POS usage after the CLC, compared with Baseline-split I and II. The CLC had a greater effect on POS usage as compared to baseline differences. Particularly, the CLC induced an increase in the usage of articles, conjunctives, and prepositions as compared to other POS categories. This increase means that the CLC changed the syntactic structures of tweets,

which is also supported by the finding that sentence length increased. Unexpectedly, the relative frequency of adverbs and adjectives did not increase after the CLC. In addition, the difference between Baseline-split I and Baseline-split II shows more variation between week 3 and week 4 as compared to week 1 and week 2. This suggests a trend in the language usage initiated by the CLC.

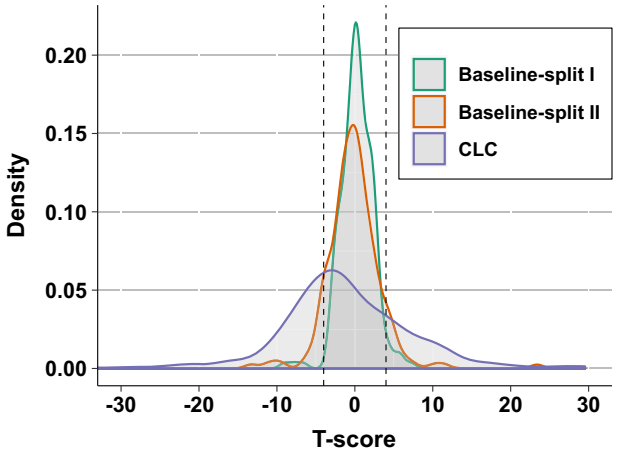

Fig. 7 *T*-score distribution of high-frequency tokens (>0.05%). The *T*-score indicates the variance in word usage; that is, the further away from zero, the greater the variance in word usage. This density distribution shows the CLC induced a larger proportion of tokens with a *T*-score lower than −4 and higher than 4, indicated by the vertical reference lines. In addition, the Baseline-split II shows an intermediate distribution between Baseline-split I and the CLC (for time-frame specifications see Fig. 1)

Discussion and conclusions

We investigated the effect of the character limit change (CLC) on the language usage in tweets. The results indicate that the CLC has, in fact, affected the language usage in tweets. The first hypothesis was supported; the pre-CLC tweets comprise relatively more textisms, such as shortenings, contractions, unconventional spellings, symbols and numerals. The second hypothesis was partially supported. As expected, the grammatical structure was affected by the CLC: post-CLC sentences are longer and comprise

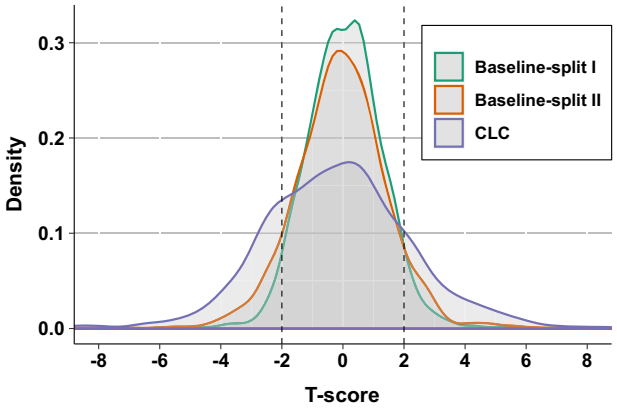

Fig. 8 *T*-score distribution of high-frequency bigrams (>0.05%). The CLC induced a larger proportion of bigrams with a *T*-score lower than −2 and higher than 2, indicated by the vertical reference lines

Table 4 Tokens that occurred relatively less frequently post-CLC and related *T*-scores (Baseline-split I; Baseline-split II; CLC)

2 (−3; −2; −36)	ff (0; −4; −9)	zeg (0; 0; −6)	omg (0; −6; −5)	ok (−1; −4; −4)
! (−1; −3; −28)	k (3; −3; −9)	& (2; −1; −6)	goeie (1; −1; −5)	ging (0; −2; −4)
? (−1; −9; −23)	hoor (0; −1; −8)	nie (−1; −7; −6)	sterkte (−2; −2; −5)	klopt (0; −3; −4)
ik (5; −13; −22)	nieuwe (0; −4; −8)	h (−2; −4; −6)	oke (1; −2; −5)	vond (−3; −1; −4)
1 (−2; −4; −21)	niks (0; 0; −8)	super (−1; −2; −6)	raar (−1; −2; −5)	s (0; 0; −4)
ja (1; −4; −17)	mn (4; −6; −8)	m (−1; −3; −6)	kapot (−1; −1; −5)	valt (−1; 1; −4)
/ (−2; −4; −16)	dacht (−1; −3; −8)	lol (2; −1; −6)	wow (0; −2; −5)	z'n (−1; 1; −4)
t (2; −3; −16)	hahaha (3; −3; −8)	kut (3; −4; −6)	jy (−3; −4; −5)	ofzo (2; −1; −4)
echt (6; −6; −15)	ie (−1; 0; −8)	ma (3; −3; −6)	1e (2; 2; −5)	lief (2; 0; −4)
v (−2; −2; −13)	plezier (−1; 0; −8)	gwn (3; −1; −6)	verjaardag (−1; −1; −5)	dood (2; 2; −4)
> (1; 1; −13)	je (−1; 0; −7)	xd (2; −3; −6)	wou (0; −1; −5)	moeder (−2; −1; −4)
heb (3; −10; −12)	al (2; −4; −7)	tis (3; −2; −6)	pfff (0; 0; −5)	ah (0; −2; −4)
3 (1; 2; −12)	me (3; −5; −7)	knap (1; 0; −6)	las (1; 0; −5)	^ (0; 0; −4)
wel (−1; −2; −11)	hij (0; −5; −7)	jaa (3; −3; −6)	oei (0; 1; −5)	p (0; −2; −4)
haha (2; −3; −11)	leuk (−1; −1; −7)	mooie (2; 3; −5)	hi (3; −1; −5)	shit (1; −1; −4)
ben (3; −4; −10)	nou (−2; 2; −7)	kijk (2; 2; −5)	eng (−1; 2; −5)	ding (1; −4; −4)
weer (3; 2; −10)	zeker (0; −2; −7)	oh (−1; −4; −5)	nr (−1; 1; −5)	wtf (5; −3; −4)
lekker (0; 0; −10)	hem (−1; −6; −7)	snap (3; 1; −5)	xx (1; −4; −5)	ha (1; 1; −4)
nee (0; −4; −10)	jou (0; −3; −7)	hou (0; −1; −5)	drama (2; 2; −5)	zn (1; −3; −4)
succes (−1; 0; −10)	kom (2; −1; −7)	top (−1; 2; −5)	jaaa (0; −2; −5)	fuck (2; −1; −4)
x (5; −4; −10)	he (3; −2; −7)	slecht (1; 3; −5)	same (2; −2; −5)	zon (3; 4; −4)
nog (1; −4; −9)	gefeliciteerd (−6; 1; −7)	idd (−2; −2; −5)	mij (4; −2; −4)	hoezo (−1; −1; −4)
zo (3; −2; −9)	+ (1; −3; −7)	heerlijk (3; −2; −5)	; (−1; 0; −4)	vet (−1; −2; −4)
was (−3; −1; −9)	ni (3; −5; −7)	gij (3; −5; −5)	n (0; −4; −4)	nice (1; −1; −4)
goed (1; 1; −9)	: (0; −4; −6)	d (−3; −1; −5)	beter (0; 1; −4)	hahah (1; −1; −4)
jij (1; 0; −9)	kan (0; −4; −6)	= (1; −4; −5)	dank (0; −1; −4)	btw (0; −2; −4)
ga (3; −2; −9)	toch (−1; −2; −6)	das (0; −2; −5)	* (3; 2; −4)	pff (0; −1; −4)
mooi (2; 1; −9)	gaat (−3; −1; −6)	7 (2; 0; −5)	m'n (2; −3; −4)	xxx (−1; −3; −4)
man (1; −1; −9)	wil (4; −1; −6)	hopelijk (−3; 0; −5)	beste (0; −1; −4)	boys (−2; 4; −4)
4 (0; 0; −9)	waarom (1; −2; −6)	ek (0; −8; −5)	leuke (0; −1; −4)	oma (−3; −1; −4)

Subset of the total 321,165 unique tokens (frequency >0.005%). The three *T*-scores represent Baseline-split I, Baseline-split II, and the CLC, respectively (see Fig. 2). Negative *T*-scores indicate a decrease in token usage and positive *T*-scores indicate an increase in token usage

Table 5 Tokens that occurred relatively more frequently post-CLC and related *T*-scores (Baseline-split I; Baseline-split II; CLC)

de (0; 10; 30)	over (2; 1; 9)	blijven (1; 0; 6)	echter (−1; 0; 5)	anderen (−1; 3; 4)
en (2; 5; 27)	onze (−3; −4; 9)	zullen (−1; −1; 6)	er (0; 1; 4)	gekregen (0; −1; 4)
van (0; 7; 22)	alle (0; 2; 9)	o.a (1; −3; 6)	daar (2; 4; 4)	zorgen (1; 1; 4)
hun (2; 7; 18)	mensen (5; 5; 8)	werden (1; 0; 6)	zelf (1; 1; 4)	ten (0; 1; 4)
het (1; 3; 17)	andere (1; 2; 8)	we (2; 2; 5)	allemaal (2; −1; 4)	enkele (−1; 3; 4)
te (0; 0; 16)	omdat (2; 1; 8)	tot (0; −2; 5)	onder (0; 1; 4)	brengen (-1; 1; 4)
, (−1; 0; 15)	bijvoorbeeld (0; 2; 8)	ons (0; 1; 5)	uw (1; −1; 4)	kort (1; −3; 4)
in (1; 3; 13)	dat (−3; 2; 7)	wij (2; 3; 5)	elkaar (−1; 2; 4)	feit (1; 2; 4)
door (2; 4; 13)	deze (1; 3; 7)	laten (0; 2; 5)	zelfs (2; −4; 4)	gebeurd (−1; −1; 4)
om (0; 1; 12)	moeten (0; 4; 7)	willen (2; 5; 5)	waren (1; 1; 4)	namelijk (1; 3; 4)
‘ (‘ (−2; 0; 12)	zoals (0; 2; 7)	eigen (0; 4; 5)	vooral (−1; 0; 4)	betreft (−2; 2; 4)
worden (2; 3; 12)	tijdens (1; −1; 7)	krijgen (2; 2; 5)	gemaakt (1; −1; 4)	voorbeeld (0; 1; 4)
met (2; 3; 11)	overigens (−2; 2; 7)	houden (−2; 6; 5)	men (0; 0; 4)	blijkt (1; −2; 4)
ze (−1; 2; 11)	hen (0; 1; 7)	mogen (1; 2; 5)	terwijl (1; −1; 4)	veranderen (1; 3; 4)
een (−3; 6; 10)	als (−1; 2; 6)	zouden (1; −1; 5)	mogelijk (−2; 1; 4)	ondanks (0; 3; 4)
die (0; 4; 10)	aan (2; −1; 6)	kleine (−1; 0; 5)	enkel (0; 0; 4)	daarmee (−1; 2; 4)
zijn (2; 3; 10)	kunnen (0; 0; 6)	plaats (−1; 1; 5)	manier (0; 1; 4)	groter (0; 0; 4)
hebben (−1; 0; 10)	maken (2; 2; 6)	zodat (1; 1; 5)	vroeger (1; 2; 4)	bepaalde (−1; 2; 4)
zich (−2; 3; 10)	want (0; −2; 6)	ter (0; 2; 5)	etc (1; 0; 4)	voldoende (−1; −2; 4)
zij (−1; −1; 10)	grote (0; 0; 6)	vervolgens (1; 1; 5)	gesprek (1; 0; 4)	waardoor (−1; 0; 4)
of (−2; −2; 9)	tussen (0; 2; 6)	waarin (0; −1; 5)	persoon (0; −1; 4)	

Subset of the total 367,896 unique tokens (frequency >0.005%). The three *T*-scores represent Baseline-split I, Baseline-split II, and the CLC, respectively (see Fig. 2). Negative *T*-scores indicate a decrease in token usage and positive *T*-scores indicate an increase in token usage

Table 6 Bigrams that occurred relatively less frequently post-CLC and related *T*-scores (Baseline-split I; Baseline-split II; CLC)

wat een (−2; 6; −10)	ook wel (0; 1; −4)	kan je (3; 0; −3)	zie ik (1; 0; −3)
ik ga (1; −4; −10)	het was (0; −2; −4)	al een (−1; −1; −3)	de enige (0; −2; −3)
ik heb (3; −7; −9)	ik had (0; −2; −4)	een mooie (0; 1; −3)	zo goed (1; 0; −3)
veel plezier (−1; −1; −9)	is toch (0; −1; −4)	nog wel (−1; −2; −3)	en dan (0; 0; −2)
ik ben (1; −3; −8)	echt een (0; −1; −4)	ik wel (0; −1; −3)	heb ik (1; −4; −2)
ik ook (2; −4; −8)	toch niet (0; −1; −4)	ga je (0; 0; −3)	en ik (0; −2; −2)
nu al (0; −4; −8)	hij is (−2; −2; −4)	de nieuwe (1; −1; −3)	ook niet (2; 0; −2)
zin in (2; −3; −8)	ik was (−2; −1; −4)	wil je (−1; 0; −3)	is niet (−1; −1; −2)
ik wil (6; −3; −7)	is nog (−1; −2; −4)	dat was (−1; 1; −3)	ook een (0; 0; −2)
heb je (0; −2; −6)	is zo (0; −2; −4)	gaan we (2; −1; −3)	ik vind (−2; 1; −2)
is echt (2; −1; −6)	op je (−2; −2; −4)	ja dat (−1; 0; −3)	vind ik (−2; −1; −2)
ik moet (1; −3; −6)	je kan (1; 0; −4)	is al (1; 0; −3)	dat hij (0; −2; −2)
dank je (0; −4; −6)	niet goed (1; 1; −4)	van mij (2; 1; −3)	ik zie (1; 1; −2)
ik dacht (−1; −1; −6)	heb een (0; −1; −4)	na een (0; 1; −3)	en nu (0; −1; −2)
jij ook (0; −1; −6)	tijd voor (2; 0; −4)	ik je (−1; −1; −3)	heel veel (2; −2; −2)
ik kan (0; −4; −5)	ja maar (−1; −2; −4)	toch wel (−2; 1; −3)	echt niet (1; 0; −2)
denk ik (1; −2; −5)	wil ik (1; 3; −4)	nog even (0; 2; −3)	moet ik (1; −3; −2)
je bent (1; 0; −5)	nog geen (2; 1; −4)	heb het (0; −1; −3)	een keer (−2; 0; −2)
is wel (−1; −1; −5)	is het (0; 0; −3)	mag ik (−1; −1; −3)	ik hoop (−2; 0; −2)
kan niet (−1; −2; −5)	dat ik (2; −6; −3)	maar wel (0; 1; −3)	maar niet (1; 1; −2)
ja ik (0; −1; −5)	als ik (1; −3; −3)	ik al (3; −1; −3)	een nieuwe (1; −1; −2)
we gaan (−1; 1; −5)	ben ik (1; −3; −3)	een goede (−1; 0; −3)	zou ik (−2; 0; −2)
ziet er (1; −1; −5)	nog niet (−1; −1; −3)	niet echt (−1; 0; −3)	had ik (2; −3; −2)
wat is (1; 1; −4)	nog een (−1; −1; −3)	ik zit (0; −2; −3)	was een (−1; 0; −2)
ben je (0; 1; −4)	ik niet (−1; −2; −3)	ik in (1; 0; −3)	een hele (−2; 0; −2)
je wel (2; −1; −4)	ook nog (−2; 0; −3)	de beste (−2; −1; −3)	die van (0; 0; −2)
ga ik (2; −1; −4)	weer een (−1; 0; −3)	volgende week (−2; 0; −3)	fijne dag (0; 0; −2)

Subset of the total 2,512,430 unique bigrams (frequency >0.015%). The three *T*-scores represent Baseline-split I, Baseline-split II, and the CLC, respectively (see Fig. 2). Negative *T*-scores indicate a decrease in relative occurrence and positive *T*-scores indicate an increase in relative occurrence

more articles, conjunctives, and prepositions than pre-CLC sentences. However, adjectives and adverbs did not increase in relative frequency. To discuss the results and implications, this section is structured as follows: first, we discuss an important insight about the results, that is, a change in the formality of language usage. After this, each of the investigated POS components are discussed separately. We conclude with possible interpretations of the results with regard to user behavior and limitations of our study.

Formality of language. The CLC seems to have brought about a qualitative change in language usage in tweets. Pre-CLC tweets contain relatively more informal language (i.e., textisms, self-

Table 7 Bigrams that occurred relatively more frequently post-CLC and related *T*-scores (Baseline-split I; Baseline-split II; CLC)

van de (0; 5; 17)	met het (1; 1; 5)	van deze (−1; 0; 4)	is de (1; −1; 2)
dat de (−2; 3; 12)	is en (0; 1; 5)	bij de (1; 2; 3)	en een (1; −1; 2)
van het (1; 2; 11)	over het (−1; 2; 5)	dat het (−2; 1; 3)	in mijn (0; −1; 2)
en de (−1; 2; 11)	aan te (1; 0; 5)	om te (0; 0; 3)	uit de (0; 0; 2)
door de (0; 2; 10)	in een (0; 1; 4)	op het (0; 0; 3)	en niet (0; 1; 2)
van een (−1; 2; 8)	voor het (2; 0; 4)	op een (0; 3; 3)	al die (0; −2; 2)
naar het (−2; −5; 8)	dat ze (0; 3; 4)	en die (−1; 0; 3)	wat je (0; 1; 2)
dat er (0; 1; 7)	te maken (0; 3; 4)	is voor (1; 2; 3)	naar een (0; 1; 2)
om de (2; 2; 7)	dan ook (−1; −1; 4)	bij een (−1; 2; 3)	en wat (0; 2; 2)
aan de (0; 1; 6)	als de (0; 1; 4)	ze niet (0; 0; 3)	dat een (−1; 2; 2)
met een (0; 1; 6)	alleen maar (1; 1; 4)	en als (0; 1; 3)	nog eens (0; 0; 2)
over de (1; 2; 6)	er zijn (0; −1; 4)	aan een (0; −1; 3)	een andere (0; 3; 2)
en het (0; 1; 6)	meer dan (0; −3; 4)	over een (−1; 0; 3)	samen met (−1; 0; 2)
mensen die (4; 3; 6)	bij het (1; −1; 4)	uit te (−1; 0; 3)	is dan (1; 0; 2)
in het (0; 3; 5)	om een (2; 0; 4)	door een (1; 2; 3)	
met de (−1; 2; 5)	op te (0; 3; 4)	in de (1; 1; 2)	
en dat (3; 3; 5)	om het (1; 1; 4)	voor de (−1; 3; 2)	

Subset of the total 2,974,471 unique bigrams (frequency >0.015%). The three *T*-scores represent Baseline-split I, Baseline-split II, and the CLC, respectively (see Fig. 2). Negative *T*-scores indicate a decrease in relative occurrence and positive *T*-scores indicate an increase in relative occurrence

Table 8 Part-Of-speech (POS) distribution

	Pre-CLC		Post-CLC		Difference	
Part-of-speech category	**Percentage**	**CI 99%**	**Percentage**	**CI 99%**	**Post-Pre**	**Relative (%)**
Adjectives	8.68	[8.65, 8.71]	8.55	[8.52, 8.57]	0.13	−1.56
Adverbs	13.3	[13.27, 13.34]	12.94	[12.90, 12.98]	0.36	−2.71
Articles	6.21	[6.18, 6.23]	6.57	[6.54, 6.59]	−0.36	5.86
Conjunctives	5.42	[5.41, 5.45]	5.68	[5.66, 5.71]	−0.26	4.72
Interjections	0.44	[0.44, 0.45]	0.38	[0.38, 0.39]	0.06	−13.32
Nouns	23.68	[23.63, 23.74]	23.64	[23.58, 23.70]	0.04	−0.19
Prepositions	9.73	[9.70, 9.76]	10.11	[10.07, 10.14]	−0.38	3.85
Pronouns	12.99	[12.96, 13.03]	12.87	[12.83, 12.90]	0.12	−0.99
Verbs	19.54	[19.49, 19.58]	19.27	[19.23, 19.32]	0.27	−1.36

All POS categories show no overlap in the 99% CI, except for nouns

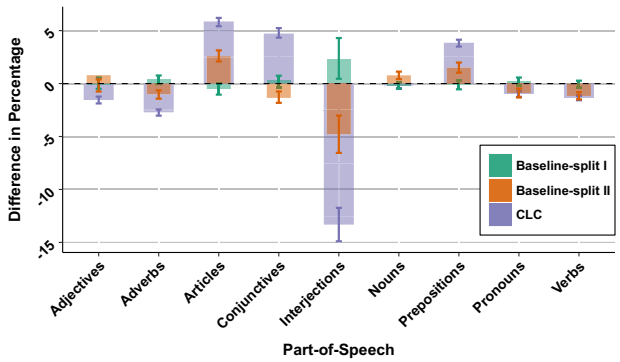

Fig. 9 Relative difference in part-of-speech usage, error bars represent 99% CIs. This bar chart shows the effect of the CLC on the part-of-speech structure of sentences as compared to Baseline-split I and Baseline-split II (for time-frame specifications see Fig. 1). The CLC induced an increase in the relative usage of articles, conjunctions, and prepositions. The relative usage of interjections decreased more than other categories and shows the highest baseline variance due to the relatively low frequency

referential pronouns, and interjection words), whereas post-CLC tweets show relatively more formal language usage. This change in formality is specifically evident in the relative frequencies of the personal pronoun *ik* (*I*) and the article word *de* (*the*), which decreased and increased, respectively. Previous n-gram research has shown that the frequencies for *ik* and *de* are indicators of informal and formal language usage (Bouma, 2015). Particularly, *ik* is used very frequently in self-referential and subjective texts such as personal social-media messages. On the other hand, *de* is used relatively more frequently in neutral and objective texts such as news articles and books. The results suggest that the CLC has led to a general change in the formality of language usage on Twitter.

POS structure. *Articles* indicate whether a noun refers to a specific entity or to an unspecified entity or class of entities (e.g., 'the house' vs., 'a house'). This information is not always essential, hence, articles can be excluded to save space or reduce the number of words, a strategy that characterizes both telegraphese and textese, (Carrington, 2004; Oosterhof and Rawoens, 2017). Articles occurred relatively more frequently after the CLC. With sufficient space, apparently, users prefer to include articles.

Conjunctions are used to link words, phrases, or clauses. The increase in conjunctions after the CLC may have multiple causes. Firstly, the relaxation of the previous restraining character limit means conjunctions are no longer 'wasting' character space, conjunctions do not necessarily have to be excluded anymore. Secondly, more available space also means there is more room for summations and subordinate clauses, thus, increasing the need for conjunctions. Another explanation for the increase in

conjunctions is the pre-CLC usage of conjunctive symbols instead of words (e.g., '/', '+', '&' as compared to '*or*', 'and').

Prepositions indicate 'where' or 'when 'an object or an individual is in relation to something else. Prepositions can describe the spatial arrangement of entities (e.g., '*The tree is in front of the house.*'). However, they are also routinely extended to depict the relations between abstract ideas, such as intentions and contrasts (e.g., '*I wear overly casual clothing to work despite the criticism from my coworkers.*'). As opposed to articles and conjunctions, most prepositions cannot be excluded without changing the conveyed meaning (e.g., '*The three is [] the house*'). Remarkably, the CLC increased preposition usage, which suggests that the prepositional information was being withheld prior to the CLC, in order to save character space. This restraint results in a truncated version of the originally intended sentence. Example (I):

Pre-CLC: 'It was a sunny beach day.'

Post-CLC: 'It was a sunny day **on** the beach, **despite** some rain **in** the morning'.

In contrast, some prepositions are omissible without changing the conveyed meaning (Rohdenburg, 2002). Example (II):

'They had difficulty [in] getting there in time.'

Both example (I) and (II) show how the relative frequency of prepositions may have increased post-CLC. However, only example (I) suggests that information was being withheld. Interestingly, the bigram analysis showed that the CLC especially increased the usage of preposition and article combinations (e.g., *by the, from the, to a)*, which appear to add non-omissible prepositional information. This finding supports the notion that information was being withheld and some sentences were obligatory truncated pre-CLC, much like example (I).

As opposed to prepositions, there was no increased usage of *adjectives* and *adverbs*. In fact, the relative usage of adjectives and adverbs *decreased* somewhat post-CLC. Adjectives and adverbs modify nouns and verbs and describe features of entities, actions, and events. For example: *'These shoes are too* (i.e., adverb) *small* (i.e., adjective).' This featural information is, perhaps, too important to be excluded from a message. When a user has to decrease word usage to remain with the character limit, it appears prepositional information is considered as expendable, whereas information related by adjectives and adverbs is regarded as indispensable. Consider the following example:

1. 'It was so nice to see my old friends and teachers from high school at the reunion.' (i.e., the original message).
2. 'Great reunion: nice to see my old high-school friends/ teachers again.' (excluding prepositions, articles, and conjunctions).
3. 'My friends and teachers from high school were at the reunion.' (excluding adjectives and adverbs).

Example 2 is clearly a more faithful rendition of the original message than example 3. Adjectives and adverbs are mainly used to describe feelings and/or opinions, which better represents the crux of a message than prepositional information. This could explain why adjective and adverb usage did not increase after the CLC.

Interjections show the largest decrease in relative frequency, see Fig. 8. The term 'interjection' is a descendent from the Latin words 'inter' and 'jacĕre' (i.e., 'to throw'). An interjection is 'thrown' between sentences and represents a sudden expression of feelings (e.g., 'Oh my!', 'Wow!', 'Haha'). Short replies mainly comprise interjections, and importantly, these interjections require very little character space. This means that the previous limit of 140 characters was already sufficient for the use of interjections. Any additional character space would therefore not be likely to affect interjection usage. This explains the relative decrease in interjection frequency compared to the other POS categories. Furthermore, the relatively low frequency of interjections also explains the higher baseline error variance as compared to the other categories.

In conclusion, the character limit change has affected language use in tweets in our sample. Tweets contained more articles, conjunctions, and prepositions, as well as relatively more formal language and relatively less informal language (i.e., textisms and interjections) after the limit change. Before the CLC, a group of users were being constrained in the conveyance of their message; post-CLC, these users obtained the character space they need. As our results show, doubling the character limit reduced the observed hindrance by a factor of ten. Therefore, the 280 characters limit appears to be much more sufficient than 140 characters to convey messages on Twitter. The new limit might appear to be a gold standard for Twitter. However, it is conceivable that, as users become more familiar with the new limit, the number of characters will increase over time. As suggested by the Baseline-split II analysis, the language usage evolves as subsequent trend of the CLC. Future research could show whether the character and language usage remains consistent or not.

Future research may also address whether the effects of the CLC in Dutch tweets are observable in other languages as well. That is, a decrease in the usage of textisms and an increase in the usage of articles, conjunctions, and prepositions. The underlying rationale being that the CLC effects are likely to be related to the function of these words and the type of information they convey, rather than the language itself. That being said, the character efficiency of the language could potentially moderate the CLC effects. Particularly, a language that is more character-efficient would be less constrained by a length limit as compared to a less character-efficient language.

An inevitable limitation of the current design is the confounding effect of natural events on the public language usage. The use of certain words can be event related. To assuage the potential impact of these confounds we removed tokens and bigrams that showed higher baseline variance as compared to the CLC-effect. However, to fully eliminate issues related to natural events, one may devise an experimental study to investigate the effect of a CLC on language usage. A CLC-dependent effect on language usage could be tested while controlling for any natural confounds (i.e., topic and event-related effects), that are bound to occur in observational studies. However, an experimental setting would reduce the ecological validity of the study. Therefore, the current study would be complementary to an experimental study.

Text-limit constraints in Tweets affect language usage, as we found in the current study. The relaxation of the character limit constraint means that writers are less likely to adapt their intended message by using strategies to compress it. Without constraints there is less need for economy of expression. The doubling of the character limit in Twitter has considerably decreased the need to compress messages. With the new limit of 280 characters, more users finally have the character space to express their thoughts. Our findings show that online language production can be affected by the character limit constraints of the medium. If necessary, language producers adapt their texts to overcome these constraints[8].

Data availability

Tweet-ids and the complete procedure are available at the Open Science Framework. It is important to note that we are not permitted to share tweets. However, we are allowed to share tweet-ids

on behalf of an academic institution and for the purpose of non-commercial research (see Developer Policy I.F.2.B. https://developer.twitter.com/en/developer-terms/policy).

Received: 27 February 2019 Accepted: 12 June 2019
Published online: 09 July 2019

Notes

1 Currently, there is much interest in algorithmic methods to define and recognize online human-behavior, such as consumer decisions, browsing activity, social-network structures, and personal interests. Twitter collects information to enhance the user experience; to show more relevant tweets, events, and people to follow, but also to enables targeted advertising (see Twitter's privacy policy; Twitter Inc, 2018). From the user's perspective, the specific implementation of personal information is unclear. That is, many of the design decisions in Twitter's software are opaque to the user. In contrast, the CLC was a transparent design decision, which directly affected the way users could interact with the Twitter environment.
2 OSF: "TCLC 1 Data Collection Pre-CLC.html" and "TCLC 2 Data Collection Post-CLC.html".
3 OSF: "tweet_ids_CLC_post.Rdata" and "tweet_ids_CLC_pre.Rdata".
4 OSF: "TCLC 3 Data Pre-Processing.html" and "TCLC 4 Data Pre-Processing 2.html".
5 OSF: "TCLC 6 Token Analysis.html" and "TCLC 7 Bigram Analysis.html".
6 OSF: "TCLC 8 Part-of-Speech Analysis.html".
7 The Dutch word 'even', which can be translated to 'for a bit' or 'just for a moment', is a commonly used filler and is often abbreviated to 'ff', which is short for "effe," a colloquial version of "even".
8 The Effect of the Twitter Character Limit Change on Language: https://osf.io/sg35a/?view_only=f360c9f624484062a43108968a4abc2b.

References

Arnholt AT, Evans B (2017) BSDA: Basic statistics and data analysis. R package version 1.2.0. https://CRAN.R-project.org/package=BSDA
Barton EL (1998) The grammar of telegraphic structures: sentential and non-sentential derivation. J Engl Linguist 26:37–67
Benoit K (2018) quanteda: Quantative analysis of textual data. R package version 0.99.22. https://doi.org/10.5281/zenodo.1004683
Bouma G (2015) N-gram frequencies for Dutch Twitter data. Computat Linguistics Netherlands 5:25–36
Buchholz S, Marsi E (2006) CoNLL-X shared task on multilingual dependency parsing. In: Màrquez L, Klein D (eds) Proceedings of the tenth conference on computational natural language learning. Association for Computational Linguistics, New York City, p 92–122
Carrington V (2004) Texts and literacies of the Shi Jinrui. Br J Sociol Educ 25:215–228
Church K, Gale W, Hanks P, Hindle D (1991) *Using statistics* in *lexical analysis*. In: Zernik Uri (ed) Lexical acquisition: exploiting on-line resources to build up a lexicon. Lawrence Erlbaum Associates, Hillsdale, p 115–164
De Jonge S, Kemp N (2012) Text-message abbreviations and language skills in high school and university students. J Res Read 35:49–68
Drouin M, Driver B (2014) Texting, textese and literacy abilities: a naturalistic study. J Res Read 37:250–267
Feinerer I, Hornik K (2017) tm: Text mining package. R package version 0.7-3. https://CRAN.R-project.org/package=tm
Frehner C (2008) Email, SMS, MMS: the linguistic creativity of asynchronous discourse in the new media age. Peter Lang, Bern
Gligorić K, Anderson A, West R (2018) How constraints affect content: the case of twitter's switch from 140 to 280 characters. In: Proceedings of the Twelfth International AAAI Conference on Web and Social Media. AAAI Press, Palo Alto
Grolemund G, Wickham H (2011) Dates and times made easy with lubridate J Stat Softw 40:1–25. http://www.jstatsoft.org/v40/i03/
Hornik K (2016) openNLP: Apache OpenNLP tools interface. R package version 0.2-6. https://CRAN.R-project.org/package=openNLP
Hornik K (2017) NLP: Natural language processing infrastructure. R package version 0.1-11. https://CRAN.R-project.org/package=NLP
Horsmann T, Erbs N, Zesch T (2015) Fast or Accurate? A Comparative Evaluation of PoS Tagging Models. In: Fisseni B, Schröder B, Zesch T (eds) Proceedings of the international conference of the German society for computational linguistics and language technology. University of Duisburg-Essen, Duisburg, p 22–30
Isserlin M (1985) On agrammatism. Cogn Neuropsychol 2:308–345
Kearney MW (2017) rtweet: collecting twitter data. R package version 0.6.0. https://cran.r-project.org/package=rtweet
Koster J (1975) Dutch as an SOV language. Linguist Anal 1:111–136
Ling R, Baron NS (2007) Text messaging and IM: Linguistic comparison of American college data. J Lang Soc Psychol 26:291–298
Lyddy F, Farina F, Hanney J, Farrell L, Kelly O'Neill N (2014) An analysis of language in university students' text messages. J Comput-Mediat Commun 19:546–561
Oosterhof A, Rawoens G (2017) Register variation and distributional patterns in article omission in Dutch headlines. Linguist Var 17:205–228
Plester B, Wood C, Joshi P (2009) Exploring the relationship between children's knowledge of text message abbreviations and school literacy outcomes. Br J Dev Psychol 27:145–161
Ratnaparkhi A (1996) A maximum entropy model for part-of-speech tagging. In: Proceedings in empirical methods in natural language processing. Association for Computational Linguistics, New Brunswick, New Jersey
Rayner K, Slattery TJ, Drieghe D, Liversedge SP (2011) Eye movements and word skipping during reading: effects of word length and predictability. J Exp Psychol: Hum Percept Perform 37(2):514–528. https://doi.org/10.1037/a0020990
R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Rohdenburg G (2002) Processing complexity and the variable use of prepositions in English. In: Cuyckens H, Radden G (eds) Perspectives on prepositions. Walter de Gruyter, Berlin, p 79–100
Rosen A, Ihara I (2017) Giving you more characters to express yourself. Blog.twitter.com. https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html
Rosen A (2017) Tweeting Made Easier. Blog.twitter.com. https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html
RStudio Team (2016) RStudio: integrated development for R. R Studio, Inc., Boston. http://www.rstudio.com/
Silge J, Robinson D (2016) tidytext: text mining and analysis using tidy data principles in R. J Open Source Softw 1(3):37
Tagliamonte SA, Denis D (2008) Linguistic ruin? LOL! Instant messaging and teen language. Am speech 83:3–34
Tesak J, Dittmann J (2009) Telegraphic style in normals and aphasics. Linguistics 29:1111–1138
Thurlow C, Brown A (2003) Generation Txt? The sociolinguistics of young people's text-messaging. Discourse Anal 1:30
Tjong Kim Sang EF (2011) Het gebruik van Twitter voor taalkundig onderzoek. TABU 39:62–72
Tjong Kim Sang EF, Van den Bosch A (2013) Dealing with big data: the case of twitter. Comput Linguist Neth 3:121–134
Twitter Inc (2018) Twitter privacy policy [PDF file]. Twitter Inc: San Francisco. https://cdn.cms-twdigitalassets.com/content/dam/legal-twitter/site-assets/privacy-page-gdpr/pdfs/PP_Q22018_April_EN.pdf
Van der Beek L, Bouma G, Malouf R, Van Noord G (2002) The Alpino dependency treebank. Lang Comput 45:8–22
Varnhagen CK, McFall GP, Pugh N, Routledge L, Sumida-MacDonald H, Kwong TE (2010) Lol: new language and spelling in instant messaging. Read Writ 23:719–733
Watson C (2017) Twitter users respond to #280characters rollout: 'All we wanted was an edit button'. The Guardian. https://www.theguardian.com/technology/2017/nov/08/twitter-users-respond-280characters-tweet-limit
Wickham H (2016) ggplot2: Elegant graphics for data analysis. Springer-Verlag, New York. http://ggplot2.org
Wickham H (2017) stringr: Simple, consistent wrappers for common string operations. R package version 1.2.0. https://CRAN.R-project.org/package=string
Wickham H, Francois R, Henry L, Müller K (2017) dplyr: A grammar of data manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr
Xie Y (2018) knitr: A general-purpose package for dynamic report generation in R. R package version 1.20
Zhu H (2018) kableExtra: construct complex table with 'kable' and pipe syntax. R package version 0.9.0. https://CRAN.R-project.org/package=kableExtra
Zwaan RA, Radvansky GA (1998) Situation models in language comprehension and memory. Psychol Bull 123:162–185

Additional information

Competing interests: The authors declare no competing interests.