# Letter replication as prosodic amplification in social media

*Susanne Fuchs*[1], *Egor Savin*[1,3], *Uwe D. Reichel*[2], *Cornelia Ebert*[1], *Manfred Krifka*[1,3]

[1]Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS) Berlin, [2]Hungarian Academy of Sciences, Budapest, [3]Humboldt-Universität zu Berlin

fuchs|savin|ebert|krifka@leibniz-zas.de, uwe.reichel@nytud.mta.hu

## Abstract

*An English blogger corpus was used to investigate letter replication as a phenomenon of prosodic emphasis in analogy to spoken language. All letters that occurred at least three times in successive order were selected. Our findings show that some letters, particularly vowels and sonorants, are favored while letters corresponding to stops were less frequently replicated. In most data the number of replications did not exceed 10 letters. Letter replication was also affected by the age of the blogger. Bloggers who were 20 years old or younger showed a greater number of replications than older bloggers. Finally, most replications were found in interjections. We suggest that social media is an interesting testbed to investigate language change, in particular the influence of spoken language on written texts.*

## Introduction

Within the last decade, written language has undergone tremendous changes. Communication has been digitalized and social media has had an increasing impact on our daily lives. These developments have gone hand in hand with technological progress (e.g. smart phones, tablets) and social media platforms (e.g. twitter, facebook, chats, whatsapp, blogs). Social media platforms provide a great opportunity to investigate the dynamics and creativity in the use of written language on social networks (e.g. Kaye et al., 2017).

Here, we will focus on a particular phenomenon of written texts: letter replication. It is sometimes even used in business, for example in advertisement of products (Figure 1).

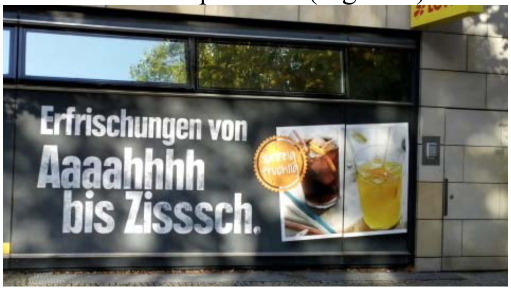

*Figure 1. Example of letter replication in advertisement at a Berlin supermarket (English translation "Refreshments from Aaaahh to Zisssch").*

Letter replication has often been associated with "word lengthening" or "vowel lengthening". Its occurrence has already been reported in twitter corpora and microblogs (Brody & Diakopoulos, 2011). Brody & Diakopoulos (2011, p. 563) suggest that word lengthening due to letter replication is a "substitute for prosodic emphasis (increased duration or change in pitch)". The phenomenon is pervasive in twitter corpora and used to emphasize words which may be important for the expression of sentiment. We may also consider letter replications as prosodic amplifiers (e.g. bbbbaaaaaaccccckkk). Amplifiers "open up a window onto a speaker's individual evaluative stance and thus express a subjective viewpoint" (Feyaerts et al. 2017, p. 486). They can also be examined in light of the recent debates about iconicity and sound symbolism. It is commonly assumed that form-meaning associations are arbitrary in language. However, small scale correspondences between form and meaning have been found in spoken language. For instance, Knoeferle et al. (2017) show that when subjects hear a syllable with a certain duration (and specific spectral properties), they judge the size of an object according to the acoustic properties, e.g. longer duration corresponds to larger objects.

We assume that the lengthening phenomena in social media have their origin in spoken language, where prosody can mark emphasis and show the speaker's individual stance. In spoken language, two lengthening aspects need to be distinguished: scope and amount. The SCOPE, i.e. the stretch of an utterance, is determined by the underlying prosodic event. This prosodic event can be a phrase boundary, a phrase-level accent, or a word stress. White (2014) gives an exhaustive literature review of these mechanisms: phrase-final lengthening extends from the nucleus of the final stressed syllable to the phrase end (e.g. Turk and Shattuck-Hufnagel, 2007). Lexical stress is marked by lengthening of the stressed syllable, particularly by the vocalic nucleus (e.g. Klatt, 1974; Oller, 1973). Finally, phrase-level accent affects the accented word. This is true predominantly for its stressed syllable but also for its boundaries (e.g., Turk and Sawusch, 1997; Turk and White, 1999).

The AMOUNT of lengthening is constrained by phoneme-intrinsic and co-intrinsic properties that are reflected in the respective average duration and variances. Several of these properties are listed in Klatt (1976), e.g. high vowels being shorter than low ones, and vowels preceding unvoiced obstruents being shorter than those preceding voiced ones. The lengthening capability of a phoneme can be expressed in terms of its elasticity (Campbell & Isard, 1991), its restoring force against being lengthened. This elasticity can be inferred from the phoneme's duration variance. Stops, for example, show a lower variance than vowels and thus can be lengthened to a lesser extent. In order to generally account for such phoneme-dependent duration behaviors, Möbius & van Santen (1996) propose a category tree for German in which phonemes are subdivided into broad classes defined in terms of phonological categories and syllable constituents. Each leaf of this tree represents a subset of phonemes as part of a certain syllable constituent. For each of these subsets a separate duration model (sum of products; van Santen, 1992) applies.

### *Research questions*

In the proposed work, we will investigate letter replication in a blogger corpus and focus on the following questions:

1.) Is letter replication only a phenomenon of vowels or also of consonants?

2.) If replications occur in consonants as well, are these only sonorants or also voiceless obstruents? Is lengthening also found in stops?

3.) How often are letters replicated? Are some cases particularly long?

4.) In which words do replications occur?

5.) How much are replications affected by social factors?

## Methodology

The Blog Authorship corpus (Schler et al., 2006) was analysed. It is a freely available corpus of blog posts from 2004 from 19,320 English speaking bloggers. It consists of roughly 140 million words and includes information about age (13-48 years with a median value of 17 years), educational background (40 different professions), sex (male vs. female), and astrological sign (12) of the bloggers.

Every blog has an xml-like structure. Only letters were selected (i.e. no numbers, emojis, symbols or replications of punctuation landmarks). All data were pre-processed to lower case. They were taken into account when they included at least 3 replications of the same letter. In some cases, triple replication of a letter may have been a typo, meaning that one should treat these cases with caution. However, many data consist of more than 3 letters (see Figure 2). For www, we checked the database and excluded all cases which were followed by punctuation marks which could refer to a weblink. Moreover, some letter replications may not constitute a word in the English lexicon, but may have specific meanings, such as “xxx” which stands e.g. for kisses (see https://www.internetslang.com/XXX-meaning-definition.asp). Hence, unlike in previous work (e.g. by Brody & Diakopoulos, 2011) we did not map the selected words with canonical word forms, because that may eliminate some data which occur only in social media, but not in the English lexicon. Data were pre-processed using Python (2.7) and graphical explorations were carried out using R (3.4.1.).

Two parameters were calculated on the basis of the dataset: Frequency of Occurrence, i.e. how often one can find a particular letter replication in the database and Number of Replication, i.e. how often a certain letter is repeated.

## Results

### *Letter replications: Where do they occur?*

Generally, replications occur for all letters (Figure 2). However, there is clearly a preference for some letter replications over others. Out of 150,147 cases with letter replication, 26.58% occur for “o” and 16.08% for “m”. Moreover, 75.05 % of all data consist of replications of “o, m, h, a, e, w”.

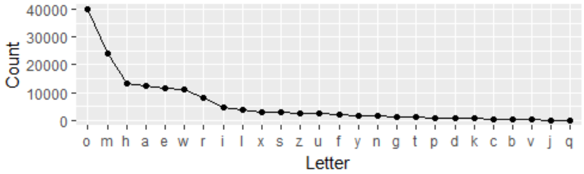

*Figure 2. Occurrence of letter replications for different letters, ordered by letters with the highest (left) to the lowest count (right).*

Except for “h”, all of these letters are sonorant and many of them are vowels. “h” may be an exception, because it often occurs after a vowel letter (e.g. Aaaahhhh in Figure 1) and may be interpreted with respect to a breathy voice quality. Figure 2 also reveals a clear disadvantage for non-continuants, i.e. stops, for letter replications. They are least frequently repeated, which is congruent with their limited lengthening in spoken language. When replications occur, they tend to be part of a general amplification with replications of most or even all letters in the word (as in bbbbaaaaaaccccckkk).

*How often are letters replicated?*

Figure 3 displays the frequency of occurrence and number of letter replication. Different colors correspond to different letters. It can be seen that most letter replications occur frequently up to 10 letters (e.g. grrrrrrrrrr). For "o" the numbers are slightly higher than for all other letters. There are also a small handful of extreme cases. For example, in one instance, "o" is repeated 4,480 times (since this is out of the range of the y-axis here, it has not been included in the figure, which aims to focus on the main observations).

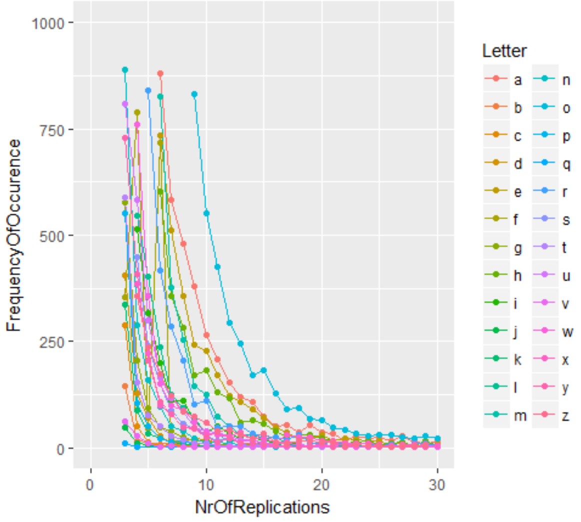

*Figure 3. Frequency of occurrence (y-axis, limited to 1000) in relation to number of replications (x-axis, limited to 30) of letter replications. Different letters are given in colors.*

*In which words do replications occur?*

The analysis so far is limited to 1000 words which were randomly selected. They were fed into a tool (open source) displaying a word cloud (https://www.jasondavies.com/wordcloud/). The results are displayed in Figure 4.

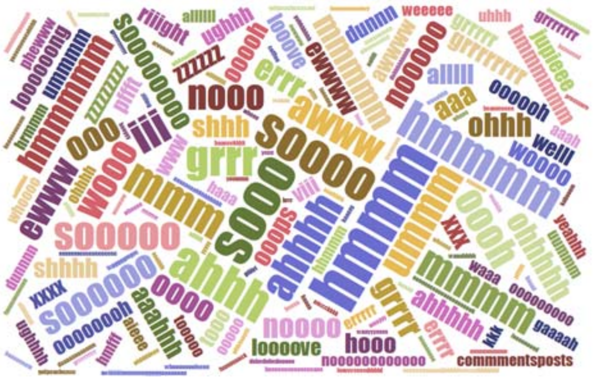

*Figure 4. Top: Word cloud plot based on 1000 randomly selected words including letter replications. The larger the font of the word, the higher its frequency (log scaled) in the selected dataset.*

The most frequent cases are interjections (e.g. hesitations such as hmmm, exclamations such as ahhh, ohhh, grrrrrr). However, a few are nouns (e.g. looove), particles (e.g. soooo), adjectives (e.g. loooooong), adverbs (e.g. tooo), pronouns (e.g. meeee), and onomatopoeias (e.g. pffft).

*Letter replication and social variables*

We also took into account the social information provided in the blogger corpus. Figure 5 shows that there is an effect of age on letter replications. It is evident that younger bloggers increasingly use replications up to the age of about 20. Between 20 and 30 years of age, the occurrence of replications decreases and then remains rather stable with older age. Furthermore, younger bloggers, particularly adolescents, produce replications with greater amplifications (soooooooooooooooooooooooooooooooooooo) than older ones (sooooooooooo). Note that for some age groups, no data are available in the corpus, which corresponds to the empty spaces in Figure 2. Males and females show comparable distributions, but the corpus contains about twice as many data for females than males.

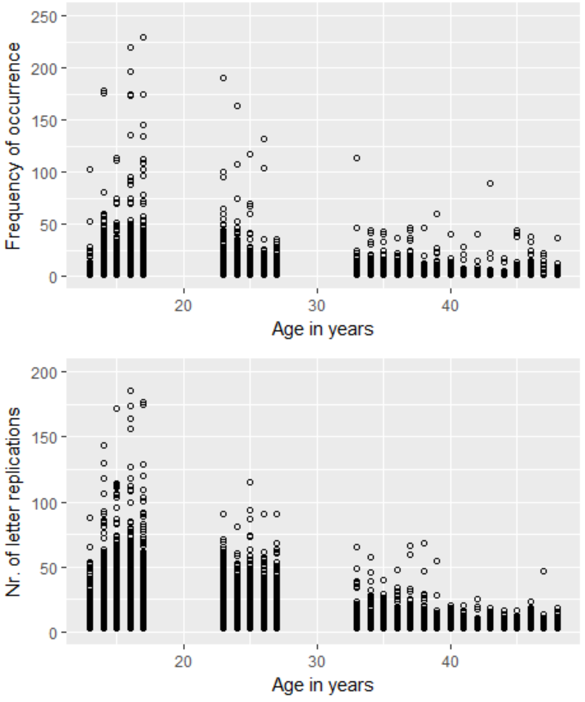

*Figure 5. Top: Frequency of occurrence of letter replications as a function of age; Bottom: Number of letter replications as a function of age (x-axis)*

The corpus also includes profession as a category. However, since profession included 40 levels, we did not further investigate these in detail. After a cursory visual inspection, the three levels categorized as "student", "education" and "unknown" used letter replications particularly frequently. We assume that there is a correlation between profession and age. These levels most likely represent younger bloggers who have not yet entered he workforce.

## Discussion and conclusion

The results of our study reveal a clear preference for the replication of certain letters and avoidance of others. Letter replication was specifically found for vowels, sonorants, and “h”. The replication of “h” occurs in most cases after a vowel (ohhh!) and we assume that it marks a lengthening of the vowel with a breathy voice. Lengthening of stops is rather limited. This result is in agreement with spoken language, where a continuous air stream is disrupted by the oral closure in stops.

The number of letter repetitions varies considerably, but we found that most data are within the range of three to ten letters, depending on the specific letter. We suppose that in spoken language, an extreme lengthening (five to ten times of the intrinsic segment duration) would not be realized, because other audio-visual cues could be utilized to express prosodic emphasis and personal stance. However, further work is needed to test this assumption.

Furthermore, we found that letter replications do not occur to the same extent in all word classes. In an initial inspection, we saw that the phenomenon was most frequent in interjections. Many of them are related to an expression of sentiment, meaning they would confirm the results by Brody & Diakopoulos (2011), even though we did not carry out a sentiment analysis.

Finally, the age of the blogger had an effect on the number of the replicated letters, with a higher number for the younger bloggers. We interpret this finding with respect to the expressiveness of the younger generation. However, it is also possible that the younger generation is more comfortable with using the specifics of digital communication due to the fact that they grew up with computer technology, while the older bloggers may still be more formal. An interesting future endeavor could be the analysis of whether emotional interjections are increasingly being replaced by emojis, which now enjoy more widespread popularity in comparison to the time at which the blogger corpus was recorded.

## Acknowledgements

This work was supported by a DFG grant to the PSIMS project (FU791/6-1). Thanks to Olivia Maky for proof reading.

## References

Brody, S., & Diakopoulos, N. (2011). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. *Proceedings of the conference on empirical methods in natural language processing* (pp. 562-570). Association for Computational Linguistics.

Campbell, W.N. & Isard, S.D. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, 19(1), 37-47.

Feyaerts, K., Oben, B., Lackner, H.K., & Papousek, I. (2017). Alignment and empathy as viewpoint phenomena: The case of amplifiers and comical hypotheticals. *Cognitive Linguistics*, 28(3), 485–509.

Kaye, L. K., Malone, S. A., & Wall, H. J. (2017). Emojis: Insights, affordances, and possibilities for psychological science. *Trends in Cognitive Sciences*, 21(2), 66-68.

Klatt, D.H. (1974).The duration of [s] in English words. *Journal of Speech, Language and Hearing Research*, 17, 51–63.

Klatt, D.H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America,* 59, 1209–1221.

Knoeferle, K., Li, J., Maggioni, E., & Spence, C. (2017). What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Scientific Reports*, 7.

Möbius, B. & van Santen, J.P.H. (1996) Modeling segmental duration in German text-to-speech synthesis. *Proceedings of. ICSLP*, Philadelphia, USA, 2395-2398.

Oller, D.K. (1973). The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America*, 54, 1235–1247.

Schler, J., Koppel, M. Argamon, S. & Pennebaker, J. (2006). Effects of age and gender on blogging. *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Turk, A.E. & Sawusch, J.R. (1997). The domain of accentual lengthening in American English. *Journal of Phonetics*, 25, 25–41.

Turk, A.E., White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics*, 27, 171–206.

Turk, A.E. & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, 35, 445–472.

van Santen, J.P.H. (1992). Contextual effects on vowel durations. *Speech Communication* 11, 513–546.

White, L. (2014). Communicative function and prosodic form in speech timing. *Speech Communication*, 63-64, 38-54.