Sobkowicz et al. *EPJ Data Science* 2013, **2**:2
http://www.epjdatascience.com/content/2/1/2

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Lognormal distributions of user post lengths in Internet discussions - a consequence of the Weber-Fechner law?

Pawel Sobkowicz[1*], Mike Thelwall[2], Kevan Buckley[2], Georgios Paltoglou[2] and Antoni Sobkowicz[3]

*Correspondence: pawelsobko@gmail.com
[1] KEN 94/140, Warsaw, Poland
Full list of author information is available at the end of the article

Abstract

The paper presents an analysis of the length of comments posted in Internet discussion fora, based on a collection of large datasets from several sources. We found that despite differences in the forum language, the discussed topics and user emotions, the comment length distributions are very regular and described by the lognormal form with a very high precision. We discuss possible origins of this regularity and the existence of a universal mechanism deciding the length of the user posts. We suggest that the observed lognormal dependence may be due to an entropy maximizing combination of two psychological factors which are perceived on a non-linear, logarithmic scale in accordance with the Weber-Fechner law, namely the time spent on post related considerations and the comment length itself. This hypothesis is supported by an experimental check of text length recognition capacity, confirming proportionality of the 'just noticeable differences' for text lengths - the basis of the Weber-Fechner law.

Introduction

Fat tailed probability distributions have been detected in many complex systems, spanning different branches of natural sciences, as well as social phenomena. Such deviations from the normal distribution usually signify some important correlations within the system. In addition to power laws, which have attracted enormous attention due to their scale-free properties (for a review see *e.g.* [1]), one of the most frequently observed forms is the lognormal distribution, found, for example, in biology, hydrology and social sciences [2–5]. Part of the reason for this generality is the origin of lognormal distribution from multiplicative random processes (forming a counterpart to additive origin of the normal distribution). In many situations (*e.g.* in population distributions or the analysis of small particle growth) it has been possible to discover directly the microscopic multiplicative mechanism driving the emergence of the lognormal distribution. Moreover, it has been shown that multiplicative processes, under slightly changed conditions, can lead to either lognormal or power law distributions [5, 6], further increasing interest in the distribution. In this work we report observations of lognormal-like properties of Internet based human communication, but postulate that their origin is not due to multiplicative process but rather to human psychology.

Human communication patterns, especially in written form, have been the subject of statistical studies for many years. Achieving machine understanding of human language has enormous practical value and a lot of research goes into automated text analysis - which is often based on statistical properties. Among the most widely known regularities are the Zipf law of word frequency distributions [7], and the distribution of sentence lengths in a given corpus of texts [8]. Williams [9] and Wake [10] have proposed that the distribution of sentence lengths is lognormal (without providing an explanation for this regularity), while Sichel criticized this approximation and used a much more complex form of the distribution [11]. The latter finding has been taken up by humanities scholars, as one of the tools allowing characterizations of style and recognition of authorship for historical texts [12–14].

Statistical studies have been applied to various forms of Internet based communications, with the goals of improving search engines efficiency, topic and attitude detection and directed advertising. In this paper we focus on a single aspect of such communication, namely on the distribution of lengths of comments that users post in various discussion fora. The comments are expressions of individual opinions and emotions, prompted by a news article or blog entry, or written as a response to another user post, thus being part of an electronic 'conversation' between two or more users. Such posts are an almost ideal medium to study the way we express our thoughts in writing, for the following reasons:

- A comment is (in most cases) a self-contained expression of thought or emotion of the author, who is in full control over the length of communication (most of the data in our study comes from electronic fora where there is no limitation on the text length).
- The comments are written with the intention to be read by other people, so they reflect an attempt at the best communication of one's feelings and thoughts to others.
- The free form of the comment allows almost total flexibility to the size of the post: from single word (or even just a single emoticon or exclamation mark) to texts comprising of tens of thousands of words (in one of our datasets we discovered a comment of almost 150,000 characters length). At the same time, the Internet allows gathering large-scale statistics. Most of the discussion fora use pure text format or text with html markup only (which can be 'cleaned' from the analysed text). In this context there is no overhead like that which is typical for word processing software, pasted-in graphics *etc.*, where the computer generated objects would obscure the user generated input.
- The social environment of the writers is not limited by education, profession or political views, providing a very broad representation of society, although perhaps over-representing groups that rely on the Internet, such as the younger generation.
- The writers in the studied fora enjoy the benefits of quasi-anonymity. This means that by using nicknames instead of real names they can be safe from immediate retribution and direct aggression offline, yet still be recognized within the discussion forum by their supporters and opponents. This removes many inhibitions present in direct communication channels or traditional forms of correspondence, where one can be personally recognized.
- Depending on the discussion forum, the comment goals range from helpful reviews (*e.g.* in hotel reservation sites or Internet bookshops) to vehement quarrels (in political blogs), so the informative and emotional content may significantly vary. This in turn may influence the social network properties linking the users and driving dynamics of post/response [15]. The studied datasets differed in such attitude and networking properties, yet the shape of the distribution of comment sizes was remarkably similar.

Previous research

The length of messages (in some cases referred to as 'size') in various forms of computer mediated human communication has been a subject of diverse studies. Especially relevant to our case are the studies of e-mails. This is because e-mails are also self-contained texts written purposefully to convey author's specific ideas or sentiments to the recipient(s). In an early study, Paxson [16] found the distribution of sizes of mail message related transfers (after subtracting the message header) to be bimodal, described by two lognormal distributions, one for the lower 80% of the data, and the other for the remaining 20%. He attributed this to the fact that some of the mails were also used to transfer files in the absence of better communication and file sharing protocols. The average size of the mail was rather small, about 1,500 bytes (characters). We recall that at that time, e-mail format was almost exclusively pure text - similar to our case of discussion posts. Another set of e-mails was analyzed by Staehle, Bolotin and Tran-Gia [17] who compared their data with older observations of Bolotin, Levy and Liu [18]. Shapes of cumulative distribution functions (CDF) for several datasets presented in [17] show quite varied behavior in the mid-size region, but they all exhibit signs of a long-tailed distribution.

These e-mail studies of data gathered in the text-based era of the 1990s can be compared with recent data obtained in large commercial environments by Karagiannis and Vojnovic [19, 20]. They found that the distribution of message sizes is almost symmetrical on a logarithmic scale - roughly lognormal in shape - but it should be noted that modern e-mails are far larger in size: the distribution spans lengths from 1,000 bytes to more than 10^7 bytes with the maximum of the distribution at about 10 kbytes. Obviously messages of megabyte size are not simple texts authored by people. The increase of size of e-mails is mainly due to prolific use of attachments, many of them containing large graphic files. The fact that the general shape is preserved suggests that a similar, lognormal distributions might apply also to the sizes of these attachments.

Indeed, such regularities were also found in studies of file systems. Douceur and Bolosky [21] found that the distribution of file sizes on a very large number of personal computers running the Windows OS follows almost ideally a lognormal distribution (with the exception of files with zero size). On the other hand, Downey [22] found that while file size distributions on a UNIX workstation follow similar lognormal patterns, there are situations where the distribution is much better described (in its fat tail) by power law, shown especially in the case of Web servers [23]. A few years later, Tanenbaum, Herder and Bos [24] confirmed such differences, noting also that for a typical UNIX workstation the shape of the file distribution has remained quite stable, but shifted to higher values. The observed duality between lognormal and power law distributions has been explained and modeled by Mitzenmacher [5], who argued that a natural mixture of lognormal distributions can yield power law behavior (see also [6]). A recent study of the distribution of sizes of files available on the Internet by Gros, Kaczor and Marković [25] has analysed a very large corpus of millions of files, with sizes ranging from 10 bytes to over 10 gigabytes. The study indicated that for some file types (for example audio and video files) the distribution of sizes follows lognormal form for large value, but for other types (images, text) it is better described by a power law. The difference was argued to originate from human neuropsychology and the maximum entropy principle.

While file sizes for complete operating system environments may be the result of specific functional requirements, there are domains where an individual file size is almost purely the result of human decisions. A very good example of such environment is provided by Wikipedia. The articles are written by large group of human users and cover a broad spectrum of topics. Here also each of the articles is a self contained message. They are mainly informative, with little emotional motivation. The article length may vary considerably, from very short to very long, although the editors provide some suggestions to the authors with respect to article length.[a] Wikipedia publishes interesting reports on the evolution of article length in various languages,[b] and although the tables cannot provide detailed information regarding the distribution, they point out the interesting effect of gradual growth in average article length. In addition to the official data, there are several studies devoted to analysis of Wikipedia entries. For example an anonymous dataset[c] shows a fat tailed shape, which is reasonably well reproduced by the lognormal distribution. In his widely cited paper, Voss [26] presents the evolution of the distribution of article sizes in German edition of Wikipedia. During four and half years the shape of the distribution changed from hugely skewed towards small sizes to a shape roughly symmetrical on a logarithmic scale (thus suggesting a lognormal distribution, but the number of data points is too small to provide a decisive measure). The process of this 'symmetrization' exhibits asymptotic behavior, with changes getting smaller when the German edition of Wikipedia became more 'mature'. In another Wikipedia study, Blumenstock used the size of an article as a measure of its quality [27, 28]. Within randomly chosen subsample of articles, the size distribution was skewed (in logarithmic scale) towards smaller sizes, suggesting deviation from the lognormal distribution. On the other hand, analysis of more extensive datasets by Serrano, Flammini and Menczer [29] confirms a very good fit of the lognormal distribution to English Wikipedia. Interestingly, in their model of formation of semantic clusters in Wikipedia, Masucci *et al.* [30] have *assumed* a lognormal distribution of sizes of 'seed' articles. Such seed articles are then mutated, the whole process aimed at reproduction of large clusters of similar articles sharing the same structure and comparable sizes, for example descriptions of geographical places or biological species. The authors do not provide links between such similarity and size distribution in the Wikipedia data corpus. Some suggestion may come from Adler *et al.* [31], where the distribution of edit sizes (changes introduced into Wikipedia articles by subsequent authors) shows a well bounded, asymmetrical distribution at logarithmic scale.

The research literature provides only a few examples of studies devoted to the medium analyzed by our paper, that is to user comments on the Internet discussion fora. For example, Mishne and Glance focus their work on the number of comments a site receives; but they provide a very rough measure of the breakdown of length of the comments, most of them bounded between 10 and 100 words, but with about 18% extending beyond 100 words [32, table 5]. Schuth, Marx and de Rijke [33], analyzing several Dutch discussion fora connected to websites of national daily newspapers, provide two distribution related figures: of the length of a comment in sentences and of the average length of a sentence in words. Both figures show roughly lognormal-like distributions, but scarcity of data does not allow a more detailed analysis. He, Caroli and Mandl [34] compare blogging site environments in China and Germany and note the observation that comments of readers (measured in sentences) are shorter in China, which they ascribe to cultural differences. In both countries, however the histogram of comment length has a skewed, fat tailed nature. Recently, Morzy [35] analyzed a discussion forum of bicycle fans at the same large newspaper web site we have used.[d] The coarse histogram of comment lengths in words presented in his work may be well described by lognormal distribution.

While electronic media provide very 'nice' subject of study, thanks to the ease of gathering and digitizing data, there are interesting examples of skewed distributions of text lengths taken from traditional, printed sources such as newspapers. Santini [36] documents lognormal distribution of the number of words per article in sample taken from Italian press. The figures shown by Leopold and Kindermann [37] contrast the distribution of lengths of Reuters news messages (showing a sharp upper bound cut-off at sizes greater than 100 words) and more complex shapes for *Frankfurter Rundschau*, where two narrow peaks, centered at about 15 and 25 words, accompany a well defined lognormal-like distribution centered at about 100 words and extending to well over 1,000. This is most likely due to the fixed organization of a newspaper layout, with short news snippets and/or main article summaries filling the visual gaps in printed layout. The broad peak would then correspond to the main articles, differing in length (but, most likely, having a lower bound cut-off, masked in the data by the additional peaks).

Fat tailed distributions in human communications are not limited to written text. A perfect example is provided by telephone conversations. In telecommunications, the talker voice volume distribution is assumed to follow a lognormal curve.[e] This assumption was a part of the standard used for planning the telecommunication networks. Call durations are also often described by lognormal distributions (*e.g.* [38]). However, a new model based on a truncated variant of log-logistic distribution has been recently proposed by Vaz de Melo *et al.* [39].

As can be seen from the above examples, despite wide range of communication modes, there are indications of general regularities in the way we express our thoughts and emotions to other people. (In this context even a volume of speech over telephone might be considered a derivative of the intention to be heard by the listener.) Unfortunately, as most of the studies focused on other issues, the data on size/length distribution is scanty, prohibiting decisive analysis and comparison of various model distribution functions and possible explanations. With this in mind, we have undertaken our study.

Datasets

In this paper we present analysis of four diverse datasets of user comments. The first is a very large corpus of data is taken from BBC discussion fora devoted to religion and ethics and to UK/world news.[f] The comments were in English. The dataset has been studied before in the context of emotions of the users and their interactions by Chmiel *et al.* [40]; here we look simply at the recorded comment length, measured in bytes or words. The raw BBC dataset contained almost 2.5 million messages in over 97,400 discussion threads, written by over 18,000 users, gathered during a period of 4 years.

The second dataset, much smaller, was gathered from several Polish news discussion fora and political blogs. This has also been originally done with the aim of analyzing emotions and opinion changes and user interaction network [15, 41]. The data contained 19,738 messages, in 57 discussion threads and from 4,718 users gathered over a period of over one year at news forum of one of most popular newspapers and a very popular political blog site.[g] Comments were in Polish. The range of emotions expressed by the users was similar to the one found in the BBC fora.

The two additional datasets were taken from MySpace discussions (taken in 2007 and 2008 in UK and US) and from YouTube comments (Thelwall, Sud and Vis [42]). These messages differed from the previous sets in the information content and formality of language. They contained a lot if informal 'netspeak': abbreviations, emoticons, misspelled words *etc.* The datasets analyzed in this work were based on 18,847 (MySpace) and 38,628 (YouTube) comments. In the latter case, less than 50% of users came from the USA and UK, with very broad international participation. We note here that from these four environments, only YouTube had a limit on allowed size of the comments, equal to 500 characters.

Message length statistics

Following previous research we have focused our analyses on the lognormal distribution, which has the Partial Distribution Function (PDF) of the form:

$$f_{LN}(L) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{L}\exp\bigl(-\bigl(\ln(L) - \mu\bigr)^2/2\sigma^2\bigr), \tag{1}$$

where L is the comment length and μ and σ are the fitting parameters. The distribution has a maximum (mode) at $c = \exp(\mu - \sigma^2)$, mean $\bar{f} = \exp(\mu + \sigma^2/2)$ and median $m = \exp(\mu)$.

Figure 1 presents the PDF distributions for the four datasets, together with the results of fits of the lognormal function, where the length of a post is measured in bytes. (The insets in Figures 1-4 show the same distributions on the log-log scale, with the data transformed to logarithmic binning, which emphasizes the behavior for large values of the comment lengths.) The fits were performed using the Levenberg-Marquardt algorithm.

Within the BBC dataset, there are a few values that 'stick out' away from the distribution significantly. These are of special origin. For example at a length of 80 bytes we observe the influence of 442 pre-formatted messages from the Editors of the forum, of the type: '`Editorial Note: This conversation has been moved from "World News" to "UK News".`' Another large deviation (798 messages) occurs at length of 115 bytes, and results from the way the gif based emoticons are converted to text in the comment harvesting algorithm. These should most likely be treated as much shorter, 'one sign' length posts. These two points, which are not user generated, are clearly visible above the rather smooth distribution. We note that the YouTube comments are limited to about 500 bytes in length. Therefore we attribute the peak at the high end of the available comment size to over representation of the comments where people used all or almost all the available space. The deviation from lognormal distribution that we observe for posts shorter that 10 characters (corresponding to one or two words) may come from the natural barrier of the shortest sensible utterance.

The maximum of the PDF function for the BBC and the Polish data is approximately 60 bytes, but due to much larger value of σ, the average length of message for the BBC forum is 434 bytes while for the Polish one it is only 257 bytes. Messages in MySpace are much shorter, the maximum of the distribution is at 25 bytes and the average is only 124 bytes; even smaller values are found in the YouTube case, 18 and 104 bytes, respectively. All the parameters are summarized in Table 1.

Figure 2 presents distributions for the BBC and the Polish fora data, with the length of the post measured in words, which are more 'human' units. They are defined here mechanically, as sets of letters separated by punctuation marks or spaces. It should be noted that because many discussion participants disregard spell-checking, situations where words are mistakenly 'glued' together are rather common. This phenomenon may cause automatic counting routines to provide results different from 'true meaning' wordcount. Another phenomenon is 'special spelling' used for emphasis, for example separating each

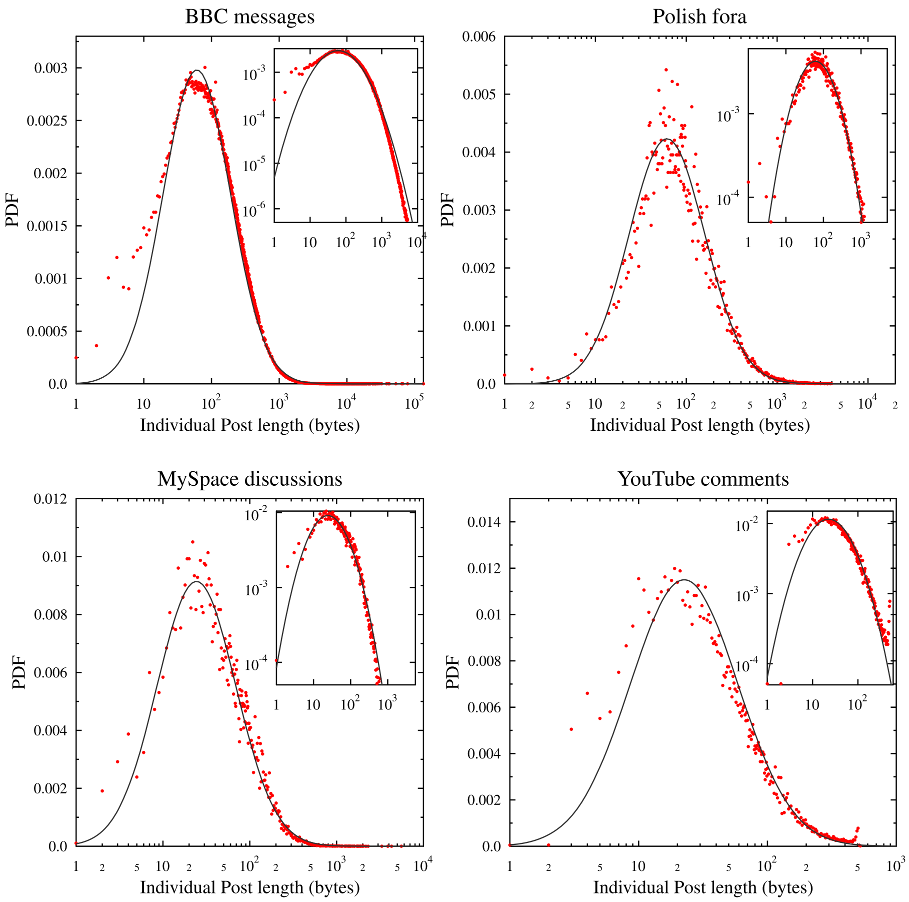


Figure 1 Post lengths in bytes (characters). Distribution of post lengths (in bytes) in various discussion fora. Top left: BBC discussions, top right political discussion fora in Poland, bottom left MySpace discussions, bottom right YouTube comments. Despite language and cultural differences the distributions are quite similar and can be well approximated with lognormal Partial Distribution Function. The parameters of the best fit are, for BBC: $\mu = 5.417$, $\sigma = 1.145$; for Polish fora: $\mu = 5.08$, $\sigma = 0.97$; for MySpace $\mu = 4.278$ and $\sigma = 1.042$ and for YouTube $\mu = 4.056$ and $\sigma = 1.086$. The insets show logbinned data on log-log scale, emphasizing the behavior for long posts.

letter of the word by spaces ‘l i k e t h i s’, which may also cause problems for automatic word counting routines. These effects are most important for the short posts, and they are the reason for relatively larger proportion of very short messages (1 to 2 words long) in the distributions for both datasets. Despite these problems we observe that the general shape of the distribution functions for long messages is very well preserved despite the change from bytes (characters) to words.

The BBC forum offered a unique opportunity to study the use of quoted text in the messages. The forum interface separated the quotes from the text written by the user, so that it was possible to select posts that openly quoted other messages and also to analyze the length of these quotes. We have separated the ‘clean’ messages, containing no graphically indicated quotations from messages that contain such insertions. Such selection applied only to the posts using special formatting, as facilitated by the forum editing mechanisms. It did not include text quoted ‘by hand’ from other sources, which we treated as original messages. It was possible to analyze this subset of messages in two ways. The

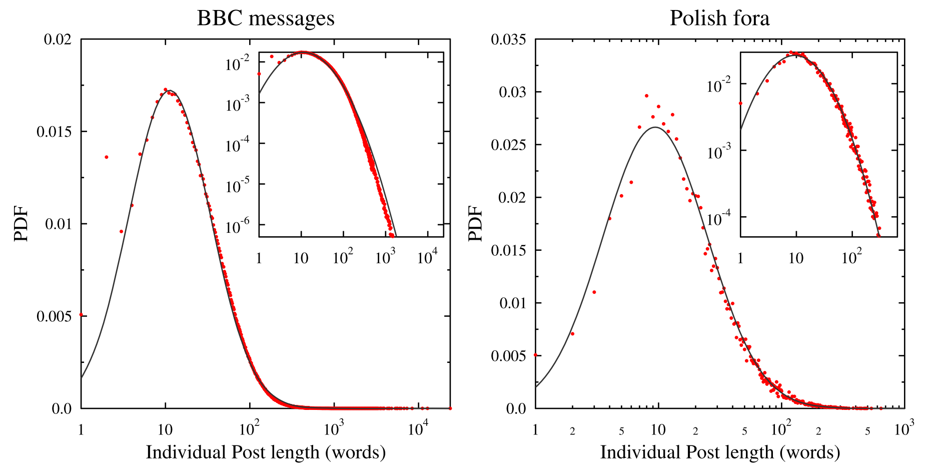


Figure 2 Post lengths in words, political fora. Distribution of post lengths (in words) in for the BBC and Polish datasets. The shape of the distribution is preserved with lognormal PDF fits parameters $\mu_W = 3.906$ and $\sigma_W = 1.26$ (BBC) and $\mu_W = 3.187$ and $\sigma_W = 0.978$ (Poland).

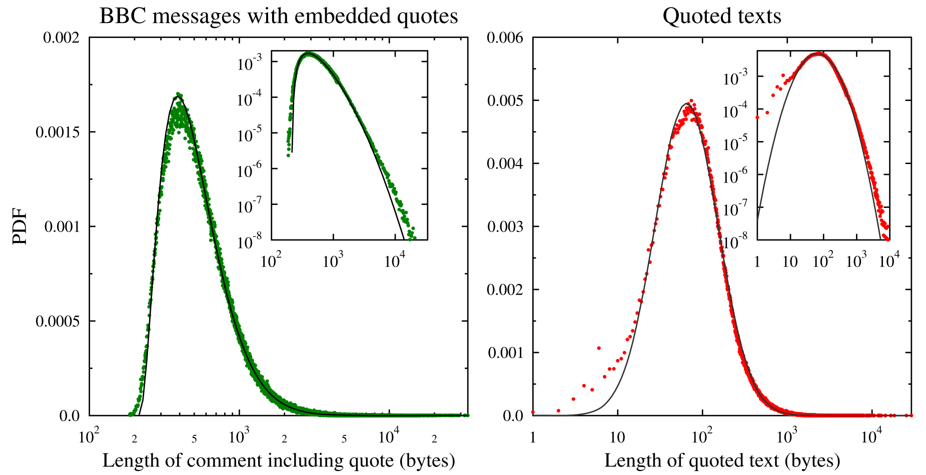


Figure 3 Distribution of lengths of posts containing embedded quotes and of the quotes. Left panel: distribution of the lengths of comments containing embedded quotes from other posts in the BBC dataset (blue dots). The distribution has a *shifted* lognormal shape $f_{LN}(x) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x-x_0}\exp(-(\ln(x-x_0)-\mu)^2/2\sigma^2)$, with $x_0 = 190$ bytes (black line). The right panel shows the distribution of the quoted text fragment lengths, well described by a the standard lognormal form, with the mean value of 196 bytes.

right panel of Figure 3 presents the distribution of the length of the quoted fragments. We note with some surprise that quoted texts have a distribution similar, but not identical, to the general distributions of the post lengths. This is interesting because intuitively the mental process deciding how much of another person's message one would cut and paste is quite different from writing one's own comment. The roles of reader and writer are different: attention is focused at selecting the part of someone else's post that might be relevant to the writer who wants to respond, support or criticize. The quote size is, in a cognitive sense, complementary to the comment being written, and the fact that the two share the same statistical properties suggests existence of some deeper mechanism, perhaps based on the user attention scope. The left panel of Figure 3 presents the PDF for

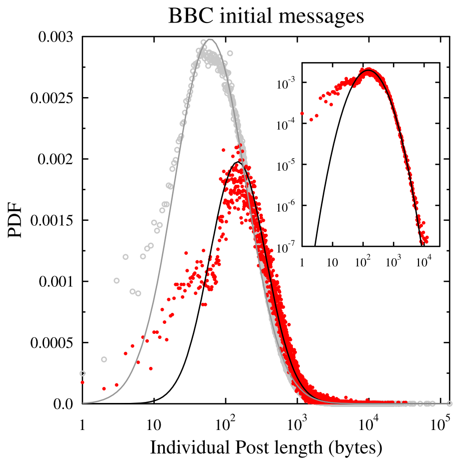


Figure 4 Distribution of lengths of posts initiating discussion thread. Distribution of the length of first messages in discussion threads for the BBC data. This distribution is important because the first messages are free from possible effects of adjusting the length of comments to previous posts. Gray circles show, for comparison, distribution of all messages in the forum (Figure 1). The initial comments show similar distribution form, but with higher average of 506 bytes and maximum of the PDF at 160 bytes. This is most likely due to the fact that initial comments carry more information and are less likely to consist of short agreement/disagreement/thanks expressions found later in the discussion threads.

Table 1 Lognormal fit parameters to observed data

	Number of messages	μ	σ	Mode	Median	Mean
BBC, all messages	2,457,586	5.417	1.145	61	225	434
BBC, first messages	97,453	5.892	0.964	143	362	576
BBC, quotes	1,123,650	4.91	0.86	65	136	196
BBC, messages with quotes (best fit lognormal distribution is shifted by 189.6 bytes)	878,382	6.045	0.839	209	422	601
Poland, all messages	19,738	5.08	0.97	63	160	257
MySpace, all messages	18,874	4.278	1.042	24	72	124
YouTube, all messages	38,628	4.056	1.086	18	58	104
BBC, all messages, words	2,457,586	3.906	1.260	10.2	49.7	109.9
Poland, all messages, words	19,738	3.187	0.978	9.3	24.2	39.1

Best fit parameters of lognormal distributions and their characteristic measures for the analysed data. Except for the last two rows, the values are in bytes (characters).

whole messages with embedded quotes. It is well described by a shifted lognormal distribution $f_{LN}(L) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{L-L_0}\exp(-(\ln(L-L_0)-\mu)^2/2\sigma^2)$; $L > L_0$, with $L_0 \approx 190$ bytes, which is very close to the mean value of length of quoted fragments (196 bytes).

Analysis of possible origins of length distribution

The regularities presented above, showing similar behavior in environments differing in the language used, cultural heritage of authors, topics of posts, presence of extended discussions between users, and emotional content of messages suggest existence of an underlying universal mechanism of a general nature. Furthermore, we think that our observations are important because Internet discussion fora provide very 'clean' examples of written communication. In contrast to formalized media (press or scientific articles, books, even web pages), which have to conform to certain standards and/or size and form limitations, the user comments are, in the studied cases, not bound by form constraints other than necessity to use written language (with the exception of limit of the comment size in YouTube, which had only a minor effect). Even this written form requirement is often circumvented, as the participants use many tricks to convey nonverbal messages: excessive use of punctuation; use of extensive capitalization of text; use of emoticons; unusual graphic forms (such as writing vertically), deliberate misspelling of names or swearwords, *etc.* The discussion posts are usually focused on a single topic (or, as happens in some cases, the posts are purposeful deviations from this topic) but there is considerable variation in the content of messages, which include informative text, humor, sarcasm, provocation and aggression [15]. The comments may be a self-contained expression of the views of the user directed to no-one in particular, or they may be a part of a conversation with other user(s). As we noted, the use of nicknames allows, at the same time, to preserve authorship recognition (in some cases discussions between users recognizing each other by nicknames continues over many separate threads, forming a true social network) and to provide safety of relative anonymity (encouraging open expression). Also, as we mentioned, free access to discussion fora means that the statistical properties are based on broad social sample.

Summarizing, we are led to conclusion that the posts are a faithful and multifaceted representation of what people want to communicate on a given subject, to general audience and/or to specific recipients. Thus the distribution of characteristic features of the comments (such as their length) may reflect our idiosyncrasies and limitations of how we want to communicate, especially using the written form.

In an attempt to explain the observed regularities we have looked for possible mechanisms driving the lognormal distribution of comment lengths. Such distributions are usually connected with random multiplicative processes, in which measured variable values are obtained by subsequent application of small multiplicative changes. In our case this would correspond to $L(P_i) = m_i L(P_{i-1})$, where m_i is a random variable centered at 1 and $L(P_i)$ is the length of post P_i. There are many situations where $m_i = (1 + \epsilon)$, where ϵ is a small random variable, corresponding to processes where subsequent modifications are relatively small and gradual, for example nanoparticle growth (*e.g.* [43, 44]) or distribution of company sizes (*e.g.* [45–47]). In these situations the actual microscopic processes leading to gradual change are well understood. The reasoning has been proposed to explain file sizes by addition/deletion of small parts of files in operating systems (*e.g.* [5]). Similar processes may be at work in clusters of Wikipedia articles sharing the same structure and related by gradual editing [30, 31].

In our case such reasoning would mean that to obtain the lognormal distribution, the posts should be variations of preceding posts or of some 'template', for example of the post starting the discussion thread. We have found that the above reasoning does not apply to our case. Separate comments are not modifications of each other. The Polish dataset has been the basis of previous research [15], which focused on opinions. Thanks to this, all the comments were read and analysed by humans. This allowed to directly check how many of the texts could be described as modifications of previous entries. As it turns out, almost all (>99%) comments were originally written (apart from in-line quotations). The posts radically differ in content from their predecessors in discussion, so the model of successive modifications of original entries, used to explain lognormal distribution of file sizes should not apply.

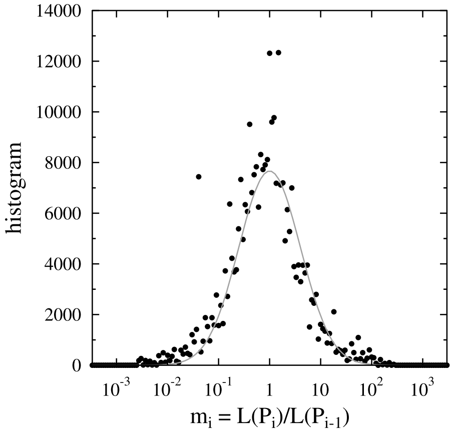


Figure 5 Ratios of lengths for consecutive messages. Histogram of natural multiplicative changes of length of consecutive messages in BBC discussions, defined as $m_i = L(P_i)/L(P_{i-1})$, on a logarithmic scale. Multiplication factors form a well defined distribution with high variation. The presence of significant number of situations when $L(P_i)/L(P_{i-1})$ is greater than 2 or smaller than 0.5 shows that even for the messages for which the content could not be analyzed, there must have been large differences in the form of new material added or large cuts. This makes the explanation of the lognormal distribution based on gradual accumulation of multiplicative changes implausible, and confirms direct analysis of the smaller datasets. Gray line is a best-fit normal distribution based on logarithm of m_i.

Due to the large size of the BBC dataset, full human analysis of the content was not possible. Spot checks of a small subset of discussions confirmed the same observation: most of the posts differ greatly from other contributions to the same thread. Interestingly, the lognormal distribution of comment sized holds also to the subset of the initial messages in the BBC dataset (Figure 4). This distribution has a higher average of 506 bytes and maximum of the PDF at 160 bytes than for the whole BBC dataset, but preserves the same functional shape. The higher values of the average and maximum of PDF are most likely due to the fact that the comments starting a thread carry more information and are cannot consist of short agreement/disagreement/thanks expressions found later in the discussion threads. As the discussion threads are separate topically and visually, there can be no 'modification' mechanism present, so the origin of the lognormal distribution remains to be explained.

For the whole BBC dataset, we have also calculated the distribution of modification factors m_i (Figure 5). The presence of large number of cases where $m_i > 2$ means that in such cases at least half of the message P_i must have been written anew, which confirms the observations from the Polish data and spot checks in the BBC set. We are therefore forced to look for another mechanism which would lead to the lognormal distribution of post sizes.

Weber-Fechner law in text length estimation and its role in comment distribution

Instead of a multiplicative mechanism, we propose here that the observed regularity is based on the Weber and Fechner laws [48, 49]. The two laws have long history in studies of human behavior. Weber law states that the just-noticeable difference (JND) in perception, differentiating between two stimuli, is proportional to the magnitude of the stimuli (usually an arithmetic mean is used). The Fechner law states that the magnitudes of the perceived intensity of external stimuli and the objective values of such stimuli are linked by the logarithmic function. It may be derived from the Weber law and for this reason both are usually grouped together. Since the mid-19th century, the Weber-Fechner law has been applied to human estimations of many physical sensations: weight, luminosity, sugar sweetness or sound loudness. It has been also applied in a variety of more complex situations, for example in estimation of a quality of service or for distances between cars on a highway [50]. The topic of the logarithmic psychological scaling has been discussed even for the basic numeral scale [51–57], unrelated to the measured/counted entity. This non-linear numerical mapping of estimates is quite general and depends on no training: it has been observed in monkeys, young children and adults with low training in mathematics and situations where we rely on intuitive estimations rather than learned numeracy. We note here that there are also many sensory phenomena for which the Weber-Fechner law was found to be inadequate, and the perceived intensity is related to the stimulus via a power-law relationship (Stevens' law [58]). The exponents for such power-law vary from much less than 1 to more than 3 [59].

We follow here the line of reasoning proposed by Gros *et al.* [25], who explained the observed regularities in file size distributions using arguments based on entropy maximizing mechanisms. In particular, the lognormal distribution is present when there are two distinct degrees of freedom in neurophysiological processes related to a particular information form. This argument has been used for audio and video files (which for large sizes follow the lognormal distribution), where the two variables were identified as time and resolution/quality. In contrast, Gros *et al.* described the text file type using a power law, associated with a single degree of freedom. We note here that this power law fit was derived mostly from the large value behavior of the distribution. The text files considered in that study ranged from 1 bytes to 10 gigabytes. Such very large files are, of course, no results of simple, focused in time human communication processes. On the other hand when one limits the considered range to below 10^5 bytes, the text file data of Gros *et al.* are quite well described by the lognormal form.

Our hypothesis, based on the idea of Gros *et al.*, explains the observed lognormal PDF shape of the originally written comments and of quoted excerpts through maximizing the entropy when the cost functions result from two independent characteristics, each following the Weber-Fechner law.

We assume that the first variable is the mental perception of the time spent on considerations related to posting of the comment or a reply to someone's post. In contrast to other forms of written communication (books, articles), the discussion posts must appear within a rather narrow time frame and are usually results of a single mental effort. In most cases the cognitive processes related to the posting activity extend well before the actual writing action. We note that time follows well documented Weber-Fechner law, valid for a broad range of values [60]. We assume here that the time spent on thinking about a post would be proportional to the post length. Such assumption is intuitively reasonable, based on the observations that short expressions of agreement, disagreement or emotions are results of quick, reflex-like activities, while longer texts require extended thinking.

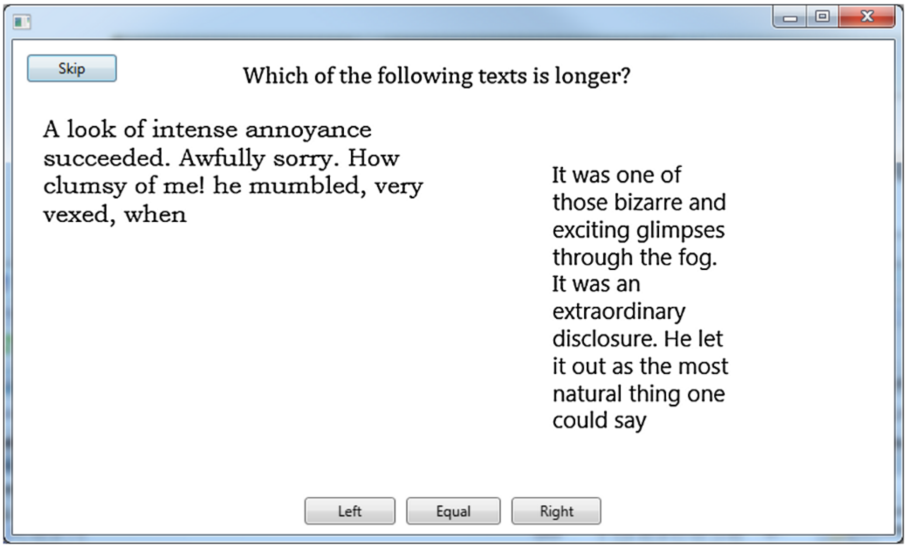
We propose that the second factor that drives the distribution is the mental perception of the text length itself. In this case, as far as we are aware of, there is no previous evidence that the objective text lengths are represented mentally on a logarithmic scale. Therefore we have decided to devise experiments that would check if the Weber-Fechner law is indeed appropriate. These experiments are described in the next sections.

To arrive at the lognormal distribution we look for general distribution function $P(L)$ that would maximize the Shannon information entropy $-\sum_L P(L)\log(P(L))$, with the constraint of the perceived cost function $c(L)$. Then the probability distribution takes the general form $P(L) \sim \exp(-\lambda c(L))$. Using the assumptions postulated above $c(L) \sim \log(T)\log(L) \sim \log(aL)\log(L) \sim (\log(a) + \log(L))\log(L)$, where the time devoted to thinking T is proportional to the length L. The resulting $P(L)$ is a general form of the lognormal distribution $P(L) \sim \exp(-\lambda_1 \log(L) - \lambda_2 \log^2(L))$.

Experiment 1

In the first experiment we have asked the participants to compare lengths of text passages, presented in a way that would deliberately obscure obvious 'graphical' cues (such as the size on page of paper or computer screen) by use of subtly differing text size, spacing, margins *etc.* The graphical user interface was designed to minimize the distractions. An example of a tester task screenshot is presented in Figure 6. The usage of different typefaces and sizes reflected a common situation found in the Internet discussion fora, where the text input window uses different font from the final display. Moreover, we wanted to avoid a possible perception shortcut via a simple area-size comparisons, forcing the participants to devote more attention to the message content, to read it.

The subject would have to judge whether a varying comparison text (of length L_C) is longer, equal or shorter than a fixed length test text (L_T), and we were looking for conditions when the user is unable to tell the difference. The test text were randomly chosen from a set with length values L_T equal to 20, 40, 50, 80, 100, 160, 200, 320, 400, 500, 640,



Which of the following texts is longer?

A look of intense annoyance succeeded. Awfully sorry. How clumsy of me! he mumbled, very vexed, when

It was one of those bizarre and exciting glimpses through the fog. It was an extraordinary disclosure. He let it out as the most natural thing one could say

Figure 6 Graphical interface of the text length recognition program - Experiment 1. The program asked the experiment subjects to compare a variety of test texts, (left side) of length L_T with another set of comparison texts (right side) of length L_C. The subjects were asked to indicate which text is longer. When a right answer was given a new comparison text, with length closer to the test one (longer or shorter than it) was substituted. If a wrong answer was given or if the subject judged the texts lengths to be equal - which indicated that the difference was below the JND threshold - a new test text was drawn from the pool. The L_T values were 20, 40, 50, 80, 100, 160, 200, 320, 400, 500, 640, 800, 1,000 and 1,280 characters, covering the spectrum from short sentences/sentence fragments to almost full page texts roughly uniformly on the logarithmic scale. All user answers were recorded in separate files for each tester, however the most interesting were those for which the subjects either considered them to be of equal length or ware unable to correctly determine the length relationships. These answers, denoted as represented length were gathered in a set of $\{L_R\}$ for each tester and for each test length L_T.

800, 1,000 and 1,280 characters, covering the spectrum from short sentences/sentence fragments to almost full page texts. These L_T values were chosen so that their distribution on a logarithmic scale would be approximately uniform. The texts were taken from classical Polish literature (native language of the studied subjects). An English language version of the program is available from the authors, as well as a Web based implementation. We performed the tests using nine testers (6 male, 3 female). The testers performed between 1,000 and 4,000 comparisons, leading to between ∼170 and ∼600 cases in which they were unable to distinguish the lengths of the strings correctly.

This experiment follows closely the original work of Weber [48] and of Cowdrick [61], where two weights were compared. We looked for the values of comparison text length for which the tested person would not be able to correctly asses in relation to the test text. Such answers comprised of two cases: the tester decided that the test and comparison texts were of the same length or he/she incorrectly assigned the length relationship (showing that he/she was unable to correctly differentiate them). We shall denote these values as represented text lengths. For each test length L_T a set of represented lengths $\{L_R\}$ was gathered for each tester. Figure 7 of the average values of represented text lengths $\langle L_R \rangle$ as functions of the test text length L_T, for each of the testers in separate panels. One can

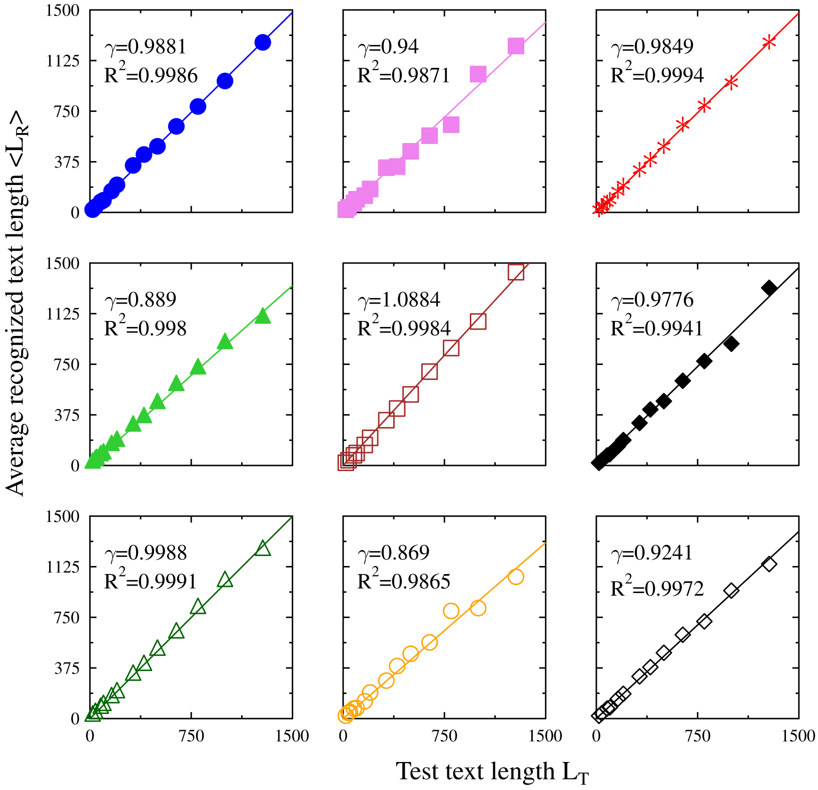


Figure 7 The average represented text length $\langle L_R \rangle$ as function test text of length L_T for nine testers in Experiment 1. For ideal perception this should be an identity $\langle L_R \rangle = L_T$. The experiments show that averages of answers is very close to this prediction and may be correctly described by a linear regression $\langle L_R \rangle = \gamma L_T$, with γ values different for each tester but in all cases close to 1. For each tester we present the value of the linear regression coefficient γ and the coefficient of determination R^2. The quality of the linear fits for $\langle L_R \rangle$ was excellent ($R^2 > 0.98$). On the other hand, the values of γ varied considerably between the testers, which indicates that the errors in answers were quite individual.

observe that the quality of the recognition is quite good. In the ideal case it would be $\langle L_R \rangle = L_T$. The observations are well described by a linear regression $\langle L_R \rangle = \gamma L_T$, with γ values different for each tester, but in all cases close to 1. Figure 8 shows the distribution of the values of γ as well as the quality of the linear fit measured by the coefficient of determination R^2. For all testers the average of the represented text length was linear with $R^2 > 0.985$, indicating extremely good linearity. The colors and symbols in Figures 7-10 denote the individual testers.

From our point of view, more interesting is the error in estimating text length, measured by the standard deviation of represented lengths, $\sigma(L_R)$. This value, calculated for various test texts of the same length L_T, would provide the JND estimation, needed to check the Weber law. Figure 8 presents the dependence of the standard deviation of the represented text lengths as function of the test text length L_T. Weber's law postulates a linear dependence $\sigma(L_R) = \beta L_T$, which, despite small number of data points, is a good fit for all the testers. The linear regression value β, linking $\sigma(L_R)$ and L_T varies between the test subjects, probably due to the level of attention during the experiment and to the individual methods used in text lengths comparisons. The quality of such assumption is presented in

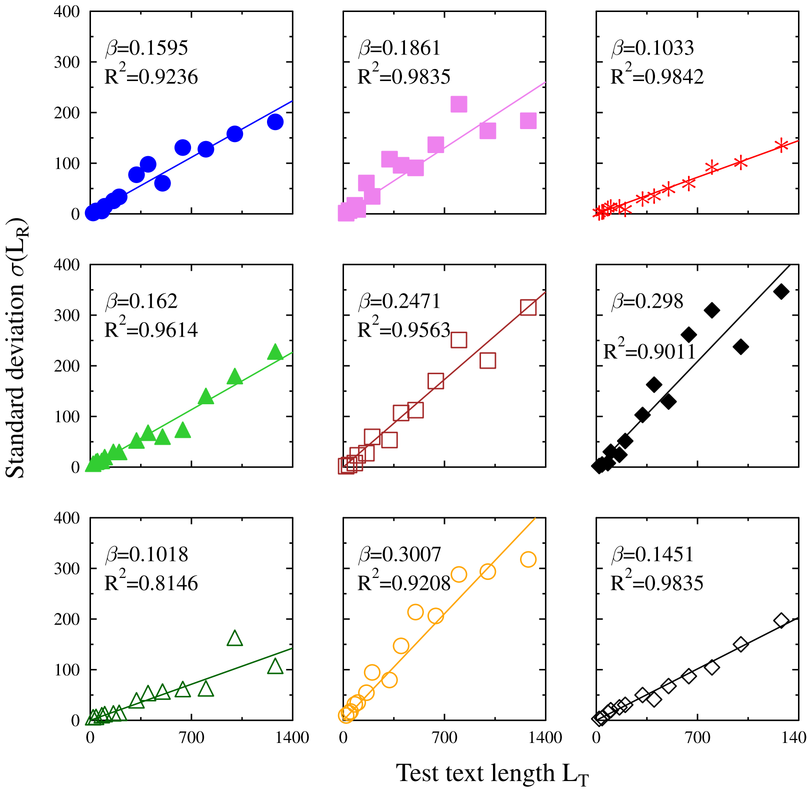
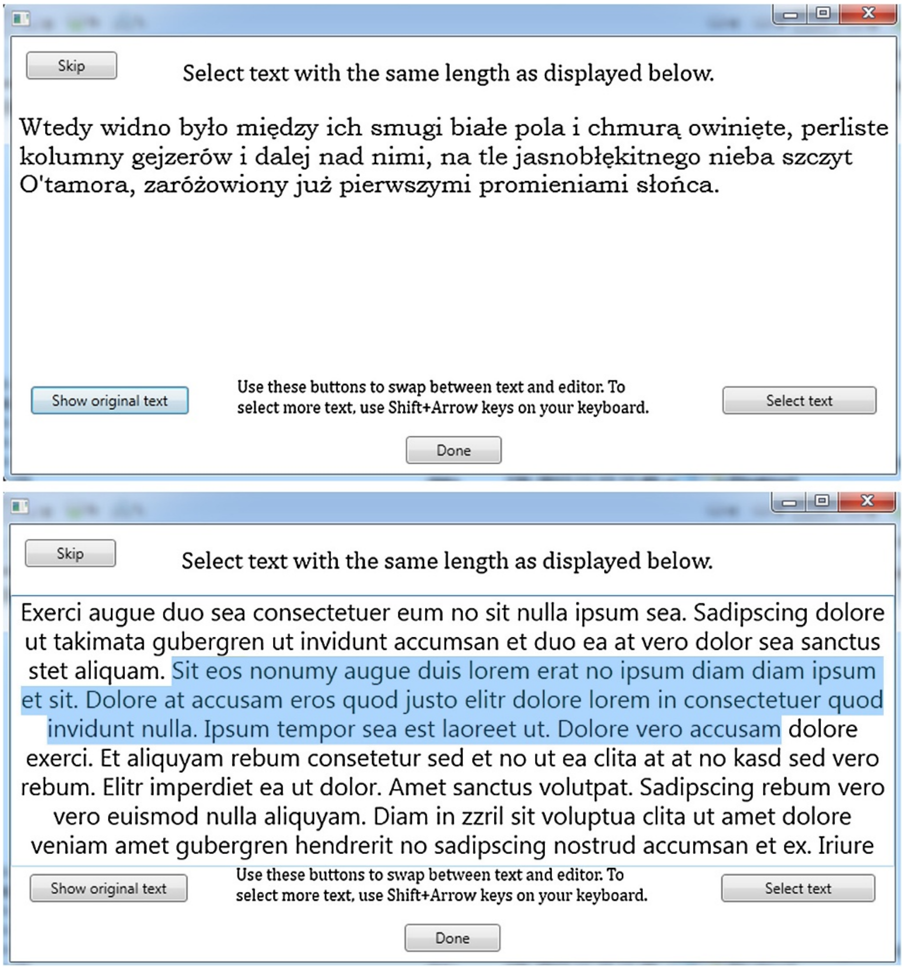


Figure 8 Linearity of JND and stimulus text lengths L_T. Comparisons of the standard deviations σ for $L_R - L_T$ for the represented lengths for each tester, serving as measure of JND, as function of the test text lengths L_T. Lines show best-fit values using linear functions corresponding to Weber law. In all individual cases the linear regression provides a good fit, $\sigma = \beta L_T$, but the accuracy of the detection, measured by β, varied strongly between the testers. For each tester we present the value of the linear regression coefficient β and the coefficient of determination R^2. The linear relationship between the standard deviation $\sigma(L_R)$ and L_T (Weber law) is well preserved with the coefficients of determination R^2 higher than 0.9 for eight of the nine subjects and above 0.8 for the remaining one.



Select text with the same length as displayed below.

Wtedy widno było między ich smugi białe pola i chmurą owinięte, perliste kolumny gejzerów i dalej nad nimi, na tle jasnobłękitnego nieba szczyt O'tamora, zaróżowiony już pierwszymi promieniami słońca.

Use these buttons to swap between text and editor. To select more text, use Shift+Arrow keys on your keyboard.

Select text with the same length as displayed below.

Exerci augue duo sea consectetuer eum no sit nulla ipsum sea. Sadipscing dolore ut takimata gubergren ut invidunt accumsan et duo ea at vero dolor sea sanctus stet aliquam. Sit eos nonumy augue duis lorem erat no ipsum diam diam ipsum et sit. Dolore at accusam eros quod justo elitr dolore lorem in consectetuer quod invidunt nulla. Ipsum tempor sea est laoreet ut. Dolore vero accusam dolore exerci. Et aliquyam rebum consetetur sed et no ut ea clita at at no kasd sed vero rebum. Elitr imperdiet ea ut dolor. Amet sanctus volutpat. Sadipscing rebum vero vero euismod nulla aliquyam. Diam in zzril sit voluptua clita ut amet dolore veniam amet gubergren hendrerit no sadipscing nostrud accumsan et ex. Iriure

Use these buttons to swap between text and editor. To select more text, use Shift+Arrow keys on your keyboard.

Figure 9 Graphical interface of the text length marking program - Experiment 2. The program required the experiment subjects to mark (using keyboard or mouse) a text fragment of the same length as the test text. The test texts were presented in the initial screen of the program (top panel). Selection was done on the selection screen (bottom panel). The users can switch between the two screens to allow corrections and calibration, but cannot see both texts at the same time. As in Experiment 1, visual clues making the purely graphical comparisons of shape and size were used. To prohibit simple content matching two different languages were used: Polish (mother language of the testers) for the test length and the 'lorem ipsum' Latin template for the markup texts. The lengths of the marked fragments denoted as corresponding length were gathered in a set of $\{L_C\}$ for each tester and for each test length L_T.

Figure 9, which shows the distribution of the proportionality constants β and the corresponding coefficient of determination R^2. For eight of nine testers the R^2 value is above 0.9 indicating good linear regression, for one tester it is somewhat greater than 0.8, a weaker, but still significant indication of linearity.

Experiment 2

In the second experiment we have asked the participants to perform a more active task. The test text (in Polish) was displayed on the computer screen (example in the upper panel of Figure 9). After familiarizing himself/herself with the test text, the subject switched to the second screen, which displayed a random fragment of a Latin text. The task was to select, using a mouse or keyboard, a fragment of the Latin text that had the same length

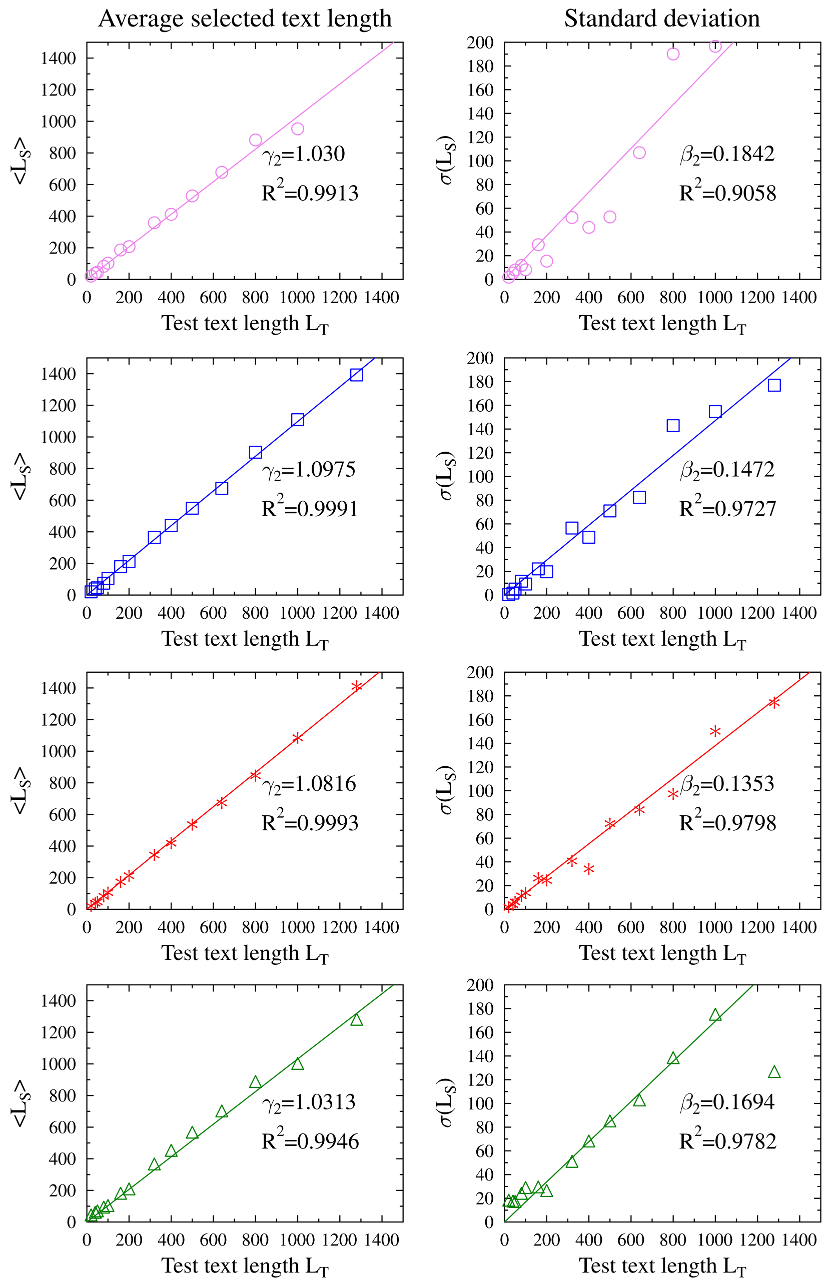


Figure 10 Results of Experiment 2. Similarly to the results of Experiment 1, the average values $\langle L_S \rangle$ (left panel) and standard deviations $\sigma(L_S)$ (right panel) as functions of L_T are plotted for four experimental subjects. The parameters of linear fits γ_2 and β_2 and the values of R^2 are shown in the figures. The origin of an outlier point for the last tester for $L_T = 1{,}280$ (with unexpectedly small error) is not understood, this point was omitted from the calculation of the linear fit.

as the original one. As before, the typefaces between the two texts were different. The use of two languages was motivated by the desire to make the matching more difficult. The subjects could switch between the two screens at will before making the final decision, but could not see the test text and the marked fragment simultaneously.

The selected passages lengths, L_S were gathered and analysed in a way similar to the Experiment 1. We have looked for the dependence of the average value $\langle L_S \rangle$ and the standard deviations $\sigma(L_S)$ on the test text lengths L_T (Figure 10). The second relationship confirms that also in the case where the experimental subject can actively choose the length of a text to fit a given value, the errors increase proportionally to the test text lengths.

Both were found to be well described by linear functions $\langle L_S \rangle = \gamma_2 L_T$ and $\sigma(L_S) = \beta_2 L_T$. In the case of the average $\langle L_S \rangle$ the linearity is excellent (with a slight tendency to overestimate the selected texts). The linearity is less pronounced for the standard deviation values; still, the coefficients of determination are better than 0.9.

Conclusions

The two simple experiments suggest the validity of the Weber's proportionality between the error and stimulus value, at least in the tested domain of the length of text between a single sentence and a single page. Due to the small number of testers the reported results should be treated as preliminary. Unfortunately, the lack of funding did not allow us to run the tests for much larger groups of participants. We are looking for institutions that would be willing to run these experiments with a larger number of testers, with the goal of obtaining the statistical distributions of the individual parameters. Despite the preliminary nature of the experimental procedure described above, it is our belief that the results support the postulated Fechner law of the logarithmic mental representation for text length. Together with the logarithmic perception of time spent at considerations related to posting a comment, these two psychological perceptions would fulfill the dual degree of freedom condition formulated by Gros, Kaczor and Marković [25] and could serve as an explanation for the observed lognormal distribution of the comment sizes.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PS has been responsible for the general concept and analysis of the data, as well as for the idea of the experimental checks of text length estimations. MT, KB and GP have gathered and prepared the BBC, YouTube and MySpace datasets. AS has gathered the Polish discussion fora datasets and implemented the software for the text length recognition testing.

Author details

[1]KEN 94/140, Warsaw, Poland. [2]School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton, WV1 1LY, UK. [3]Faculty of Mechatronics, Warsaw University of Technology, Św. Andrzeja Boboli 8, Warsaw, 05-525, Poland.

Acknowledgements

The second, third and fourth authors were supported by a European Union grant from the 7th Framework Programme, Theme 3: Science of complex systems for socially intelligent ICT as part of the CyberEmotions project (contract 231323). The first author is grateful to Mr Ksawery Stojda, for fruitful discussions and suggestions. We would also like to thank the anonymous referee for directing us to the Gros *et al.* [25] paper, which provided significant insight into the origin of the lognormal distribution in our data.

Endnotes

[a] http://en.wikipedia.org/wiki/Wikipedia:Article_size.

[b] http://stats.wikimedia.org/EN.

[c] http://en.wikipedia.org/wiki/File:English_Wikipedia_Article_Length_Statistics.png, the Wikipedia user Nakon who posted the figure does not reveal his real-life identity.

[d] www.gazeta.pl.

[e] ANS T1.523-2001 standard Telecom Glossary 2000, http://www.atis.org/glossary/definition.aspx?id=4017.

[f] www.bbc.co.uk/dna/mbreligion, www.bbc.co.uk/dna/mbfivelive/F2148565, www.bbc.co.uk/dna/mbfivelive/F2148564.

[g] www.gazeta.pl, kataryna.blox.pl.

Received: 20 August 2012 Accepted: 13 February 2013 Published: 18 February 2013

References

1. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. Contemp Phys 46:323-351

2. Aitchison J, Brown J (1963) The lognormal distribution. Cambridge University Press, Cambridge
3. Limpert E, Stahel W, Abbt M (2001) Log-normal distributions across the sciences: keys and clues. Bioscience 51(5):341-352
4. Mitzenmacher M (2004) A brief history of generative models for power law and lognormal distributions. Internet Math 1(2):226-251
5. Mitzenmacher M (2004) Dynamic models for file sizes and double Pareto distributions. Internet Math 1(3):305-333
6. Reed W, Jorgensen M (2004) The double Pareto-lognormal distribution - a new parametric model for size distributions. Commun Stat, Theory Methods 33(8):1733-1754
7. Zipf GK (1935) The psycho-biology of language: an introduction to dynamic philology. Houghton Mifflin, Boston
8. Yule G (1939) On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. Biometrika 30(3-4):363
9. Williams CB (1940) A note on the statistical analysis of sentence-length as a criterion of literary style. Biometrika 31(3-4):356
10. Wake W (1957) Sentence-length distributions of Greek authors. J R Stat Soc A, General 120(3):331-346
11. Sichel H (1974) On a distribution representing sentence-length in written prose. J R Stat Soc A, General 137:25-34
12. Montemurro M, Zanette D (2002) New perspectives on Zipf's law in linguistics: from single texts to large corpora. Glottometrics 4:87-99
13. Zanette D, Montemurro M (2005) Dynamics of text generation with realistic Zipf's distribution. J Quant Linguist 12:29-40
14. Sigurd B, Eeg-Olofsson M, Van Weijer J (2004) Word length, sentence length and frequency - Zipf revisited. Stud Linguist 58:37-52
15. Sobkowicz P, Sobkowicz A (2010) Dynamics of hate based Internet user networks. Eur Phys J B 73(4):633-643
16. Paxson V (1994) Empirically derived analytic models of wide-area TCP connections. IEEE/ACM Trans Netw 2(4):316-336
17. Staehle D, Leibnitz K, Tran-Gia P (2000) Source traffic modeling of wireless applications. Technical report No. 261, Universität Würzburg, Institut für Informatik
18. Bolotin V, Levy Y, Liu D (1999) Characterizing data connection and messages by mixtures of distributions on logarithmic scale. In: ITC-16: international teletraffic congress, pp 887-894
19. Karagiannis T, Vojnovic M (2008) Email information flow in large-scale enterprises. Technical report, Microsoft research. ftp://ftp.research.microsoft.com/pub/TR/TR-2008-76.pdf
20. Karagiannis T, Vojnovic M (2009) Behavioral profiles for advanced email features. In: Proceedings of the 18th international conference on World Wide Web, pp 711-720
21. Douceur J, Bolosky W (1999) A large-scale study of file-system contents. ACM SIGMETRICS Perform Eval Rev 27:59-70
22. Downey A (2001) The structural cause of file size distributions. In: Proceedings of ninth international symposium on modeling, analysis and simulation of computer and telecommunication systems, pp 361-370
23. Arlitt M, Williamson C (1996) Web server workload characterization: the search for invariants. ACM SIGMETRICS Perform Eval Rev 24:126-137
24. Tanenbaum A, Herder J, Bos H (2006) File size distribution on UNIX systems: then and now. ACM SIGOPS Oper Syst Rev 40:100-104
25. Gros C, Kaczor G, Markovic D (2012) Neuropsychological constraints to human data production on a global scale. Eur Phys J B 85:28
26. Voss J (2005) Measuring Wikipedia. In: International conference of the international society for scientometrics and informetrics, pp 221-231
27. Blumenstock J (2008) Automatically assessing the quality of Wikipedia articles. Technical report, School of Information, UC Berkeley. http://escholarship.org/uc/item/18s3z11b
28. Blumenstock J (2008) Size matters: word count as a measure of quality on Wikipedia. In: Proceedings of the 17th international conference on World Wide Web, pp 1095-1096
29. Serrano M, Flammini A, Menczer F (2009) Modeling statistical properties of written text. PLoS ONE 4(4):e5372
30. Masucci AP, Kalampokis A, Eguiluz VM, Hernández-Garcia E (2011) Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. PLoS ONE 6(2):e17333
31. Adler B, de Alfaro L, Pye I, Raman V (2008) Measuring author contributions to the Wikipedia. In: Proceedings of the 4th international symposium on Wikis, pp 1-10
32. Mishne G, Glance N (2006) Leave a reply: an analysis of weblog comments. In: Third annual workshop on the weblogging ecosystem
33. Schuth A, Marx M, de Rijke M (2007) Extracting the discussion structure in comments on news-articles. In: Proceedings of the 9th annual ACM international workshop on web information and data management. ACM, New York, pp 97-104
34. He Y, Caroli F, Mandl T (2007) The Chinese and the German blogosphere: an empirical and comparative analysis. In: Mensch & Computer 2007: Konferenz für interaktive und kooperative Medien, München, pp 149-158
35. Morzy M (2009) On mining and social role discovery in Internet forums. In: International workshop on social informatics, SOCINFO '09. IEEE Press, New York, pp 74-79
36. Santini M (2001) Text typology and statistics. Explorations in Italian press subgenres. Ital J Linguist 13:339-374
37. Leopold E, Kindermann J (2002) Text categorization with support vector machines. How to represent texts in input space? Mach Learn 46:423-444
38. Willkomm D, Machiraju S, Bolot J, Wolisz A (2008) Primary users in cellular networks: a large-scale measurement study. In: New frontiers in dynamic spectrum access networks, 2008. DySPAN 2008. 3rd IEEE symposium on, pp 1-11
39. Vaz de Melo P, Akoglu L, Faloutsos C, Loureiro A (2010) Surprising patterns for the call duration distribution of mobile phone users. In: Balcázar J, Bonchi F, Gionis A, Sebag M (eds) Machine learning and knowledge discovery in databases. Lecture notes in computer science, vol 6323. Springer, Berlin, pp 354-369
40. Chmiel A, Sobkowicz P, Sienkiewicz J, Paltoglou G, Buckley K, Thelwall M, Holyst J (2011) Negative emotions boost users activity at BBC forum. Physica A 390(16):2936-2944
41. Sobkowicz P, Sobkowicz A (2012) Properties of social network in an Internet political discussion forum. Adv Complex Syst. doi:10.1142/S0219525912500622

42. Thelwall M, Sud P, Vis F (2012) Commenting on YouTube videos: from Guatemalan rock to El Big Bang. J Am Soc Inf Sci Technol 63(3):616-629
43. Söderlund J, Kiss L, Niklasson G, Granqvist C (1998) Lognormal size distributions in particle growth processes without coagulation. Phys Rev Lett 80(11):2386-2388
44. Espiau de Lamaëstre R, Bernas H (2006) Significance of lognormal nanocrystal size distributions. Phys Rev B 73(12):125317
45. Gibrat R (1931) Les inégalités économiques
46. Stanley M, Buldyrev S, Havlin S, Mantegna R, Salinger M, Stanley HE (1995) Zipf plots and the size distribution of firms. Econ Lett 49(4):453-457
47. Sutton J (1997) Gibrat's legacy. J Econ Lit 35:40-59
48. Weber EH (1846) Der Tastsinn und das Gemeingefühl. In: Wagner R (ed) Handwörtebuch der Physiologie. Vieweg, Braunschweig, pp 481-588
49. Fechner GT (1860) Elemente der Psychophysik, vol 3. Breitkopf & Härtel, Wiesbaden
50. Murata T (1988) A basic study on safe car-following distance - report of the national research institute of police science. Research on traffic safety and regulation. Publication of National Police Agency, Japan 29:17-23
51. Dehaene S (2003) The neural basis of the Weber-Fechner law: a logarithmic mental number line. Trends Cogn Sci 7(4):145-147
52. Siegler R, Opfer J (2003) The development of numerical estimation. Psychol Sci 14(3):237
53. Dehaene S, Izard V, Spelke E, Pica P (2008) Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. Science 320(5880):1217-1220
54. Cantlon J, Cordes S, Libertus M, Brannon E (2009) Comment on "Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures". Science 323(5910):38
55. Dehaene S, Izard V, Pica P, Spelke E (2009) Response to comment on "Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures". Science 323(5910):38
56. Izard V, Dehaene S (2008) Calibrating the mental number line. Cognition 106(3):1221-1247
57. Merten K, Nieder A (2009) Compressed scaling of abstract numerosity representations in adult humans and monkeys. J Cogn Neurosci 21(2):333-346
58. Stevens S (1957) On the psychophysical law. Psychol Rev 64(3):153
59. Stevens S (1975) Psychophysics: introduction to its perceptual, neural, and social prospects. Transaction Publishers, Picataway
60. Lewis P, Miall R, Lewis P, Miall R (2009) The precision of temporal judgment: milliseconds, many minutes, and beyond. Philos Trans R Soc Lond B, Biol Sci 364(1525):1897-1905
61. Cowdrick M (1917) The Weber-Fechner law and Sanford's weight experiment. Am J Psychol 28(4):585-588

doi:10.1140/epjds14
Cite this article as: Sobkowicz et al.: **Lognormal distributions of user post lengths in Internet discussions - a consequence of the Weber-Fechner law?.** *EPJ Data Science* 2013 **2**:2.