Meaningfulness and unit of Zipf's law: evidence from danmu comments

Yihan Zhou
Department of East Asian Languages and Cultures
University of Illinois at Urbana-Champaign
707 S Mathews Ave, Urbana, IL, USA 61801
yzhou114@illinois.edu

Abstract

Zipf's law is a succinct yet powerful mathematical law in linguistics. However, the meaningfulness and units of the law have remained controversial. The current study uses online video comments call "danmu comment" to investigate these two questions. The results are consistent with previous studies arguing Zipf's law is subject to topical coherence. Specifically, it is found that danmu comments sampled from a single video follow Zipf's law better than danmu comments sampled from a collection of videos. The results also suggest the existence of multiple units of Zipf's law. When different units including words, n-grams, and danmu comments are compared, both words and danmu comments obey Zipf's law and words may be a better fit. The issues of combined n-grams in the literature are also discussed.

1 Introduction

1.1 Zipf's law

Zipf's law is an important empirical law describing the statistical properties of many natural phenomena. The law states that the frequency of a word in a given corpus has an inverse proportion with its frequency rank (Zipf, 1949). Ideally, the word that ranks first will be twice as frequent as the word that ranks second and so forth. This quantitative relation is captured in the following equation where f represents word frequency, r stands for the frequency rank, and C is a constant.

$$f = \frac{C}{r} \tag{1}$$

Mandelbrot (1953) proposed a refinement of Zipf's original equation by adding two constants m and β. m is the shifted rank and β is the exponent which is estimated to be 1. The following equation shows Mandelbrot's revision:

$$f = \frac{C}{(r+m)^{\beta}} \tag{2}$$

Zipf's law is quite common in natural languages and language-related phenomena. It was found that the translated versions of the Holy Bible in one hundred natural languages approximately follow Zipf's law (Mehri and Jamaati, 2017). Zipf's law also holds for artificial languages such as Esperanto, programming languages such as Python and UNIX (Chen, 1991; Manaris et al., 2006; Sano et al., 2012).

Beyond languages, Zipf's law also has a wide coverage in physical, biological, and behavioral phenomena. Examples include city sizes, webpage visits, scientific citation numbers, earthquake magnitudes (Cunha et al., 1995; Gabaix, 1999; Redner, 1998; Newman, 2005).

1.2 Remaining questions in Zipf's law

Zipf's law has been proposed for more than 70 years, yet a central question still persists: why the complex language production processes should conform to a mathematically concise equation (Piantadosi, 2014). While many studies have successfully demonstrated Zip's law in languages, very few explained the underlying cause of word frequency distribution (Manin, 2008). It is thus crucial for any explanation of Zipf's law to make new predictions and have their assumptions tested with more data (Piantadosi, 2014).

Another question that has been little addressed is the unit in Zipf's law (Corral et al., 2015). In the literature, the majority of studies have used word as the frequency unit to derive Zipf's law (Corral et al., 2015). However, word may not always be the right unit since the meaningful components in languages are a combination of words and phrases (Williams et al., 2015). Moreover, word as an umbrella term can be difficult to define linguistically (Dixon and Aikhenvald, 2002). Even if we get by with words, empirical data show that other units such as phrases and combined n-grams sometimes fit Zipf's law better than words (Ha et al., 2009; Williams et al., 2015). Therefore, it is important to compare how different units obey Zipf's law before taking word for granted as the default unit.

1.3 Goal of the current study

The current study uses danmu comments as data to explore the meaningfulness and unit of Zipf's law. It aims to answer three questions: 1. Do danmu comments follow Zipf's law? 2. Is there an optimal unit of Zipfian distribution in danmu comments? 3. What does the distribution of danmu comments imply for the meaningfulness of Zipf's law?

The study can contribute to research on Zipf's law in three ways. First, it extends Zipf's law to new data. Although it is tempting to assume that Zipf's law is universal, Li (2002) advocated that we should examine the data first. In that case, we can identify new data which follow Zipf's law and reject false data that are assumed to exhibit Zipf's law. Second, it can advance our understanding of Zipf's law in internet language. Previous studies have examined how search terms in searching engine and tags in online blogs exhibit Zipf's law (Chau et al., 2009; Liu, 2008). However, these items are not very different from words in regular texts in terms of length and composition. In comparison, danmu comments have features that are not commonly found in normal texts such as a large amount of code-mixing and neologisms. Finally, comparing how different units follow Zipf's law may provide indirect evidence to the cause and meaningfulness of Zipf's law.

1.4 Danmu comments

Danmu comments, or danmaku in Japanese, is an emerging type of commentary system for online videos (Wu et al., 2019). Danmu comments first appeared in Niconico, a Japanese video sharing website and spread to China afterwards (Yao et al., 2017). Danmu comments are scrolling anonymous comments on the screen that allow participants to express feelings or opinions while watching a video (Bai et al., 2019). When danmu comments become dense, they can cover the entire screen and create a visual impression resembling the artillery barrage in warfare. Therefore, this type of comments acquires the name "danmu" which literally means "barrage" in Chinese (Chen et al., 2015).

There are three characteristics of danmu comments worth pointing out. First, danmu comments are comprised of diverse symbols, including linguistic symbols, digits, punctuation, emoji, etc. (Li, 2018). Second, danmu comments often employed homophones called mishearing or soramimi (Nakajima, 2019). Some danmu comments sound similar to what is said in the video, but convey different meanings. Third, danmu comments have independent meanings from the video such that users can be as much interested in the danmu comments as in the video itself (Nakajima, 2019). Some users just watch a video for the sake of danmu comments.

2 Previous work

2.1 Meaningfulness of Zipf's law

Despite its seeming omnipresence, the origin of Zipf's law remains a controversy (Cancho and Solé, 2003). In particular, whether Zipf's law is meaningful has been heatedly debated. In the literature, there are mainly four accounts of Zipf's law in natural languages: random account, stochastic account, communication account, and semantic account (Piantadosi, 2014).

The first account is the random account. This view demonstrates mathematically that random texts can exhibit Zipf's law and concludes Zipf's law is purely a statistical phenomenon without linguistic meanings (Mandelbrot and Mandelbrot, 1982; Li, 1992). For example, Li (2002) created random texts by inserting M + 1 symbols in each position with one of the symbols being the word boundary. The random "words" were then extracted based on the word boundary symbol. Both mathematical proof and numerical simulation showed that the random "words" follow Zipf's law. Similarly, random sequences consist of symbols from a set of M symbols with equal probability are also shown to conform to Zipf's law (Zörnig, 2015).

The second view draws on a stochastic model. Simon (1955) postulated that a power-law distribution will take shape if new elements grow at a constant rate and old elements reoccur at a rate proportional to their probability in all elements that have appeared. Similarly, Barabási and Albert (1999) proposed that growth and preferential attachment are the origin of scale-free power-law distribution. On the one hand, new vertices are added as the network grows. On the other hand, new vertices prefer to attach to vertices that already have many connections. Removing either factor will eliminate the scale-free features in the network.

The third view argues the origin of Zipf's law is the results of optimal communication. Zipf (1949) proposed least effort is the fundamental principle governing all human actions. This principle entails two types of economy: the speaker's economy which prefers to express all meanings with one word and the auditor's economy that favors a one-to-one mapping between meanings and word forms. Ultimately, Zipf's law is a vocabulary balance between these two conflicting forces. Cancho et al. (2003) implemented Zipf's idea of least effort with an energy function defined as the sum of the effort for the hearer and the effort for the speaker. The model showed that Zipf's law is the compromise between the needs of the both the hearer and the speaker.

Finally, the semantic view presumes that word frequency is determined by semantics. It is argued that word meanings tend to expand and people are reluctant to use too many synonyms. Under the influence of these two forces, words meanings develop into a semantic space with multiple layers, which give rise to Zipf's law (Manin, 2008).

2.2 Units in Zipf's law

Studies on Zipf's law in Indo-European languages have predominately used word as the unit. However, as mentioned before, the concept of word can be ambiguous. More importantly, empirical evidence shows that other linguistic units may also fit Zipf's law.

Kwapień and Drożdż (2012) studied the distribution of words and lemmas (i.e. the dictionary form of words) in one English text and one Polish text. It was found that words and lemmas have similar distribution in the English text, but not in the Polish text. It was also pointed out that Zipfian-like scaling covers wider range in words than in lemmas.

In a follow-up study, Corral et al. (2015) conducted a large-scale comparison on how words and lemmas follow Zipf's law in single-authored texts of 4 different languages. These languages range from morphologically poor language to morphologically rich language (i.e. English, Spanish, French, and Finnish). The authors found that Zipf's law holds for both word and lemmas.

Williams et al. (2015) compared different linguistic units with three kinds of text partition: (1) no whitespace serves as the word boundary and clauses remain clauses (2) each whitespace has 50% chance of being the word boundary and clauses are cut into phrases of one or more words (3) every whitespace is treated as the word boundary and clauses are segmented into words. The results showed that both words and phrases yielded β (i.e. exponent of Zipf's law) close to 1, but phrases ($\beta = 0.95$) may be a better fit than words ($\beta = 1.15$).

Ha et al. (2009) extend Zipf's law to n-grams in English, Latin, and Irish with large corpora. It was found that word frequency in English follows Zipf's law only up to the rank of 5000 while Latin and Irish words follow Zipf's law till the rank of 30000. In addition, none of the individual n-grams (i.e. bigrams, trigrams, 4-grams, 5-grams) in the three languages fit the Zipf's curve. However, when all the n-grams from unigrams to 5-grams are combined, the data in the three languages all become close to Zipf's curve for almost all ranks. The study suggested that combined n-grams are a better fit for Zipf's law than words.

For languages like Chinese and Tibetan, there are no whitespaces to mark the word boundary. As a result, multiple linguistic units exist in the written corpora. For example, apart from word, Chinese also has a conventional linguistic unit called "character". Character is the basic unit in Chinese writing system and one character often corresponds to one morpheme and one syllable (Deng et al., 2014).

Guan et al. (1999) compared Zipf's law in three linguistic units including characters, words, and bigrams and concluded that Zifp's law applies to all three units. Another study also showed that Chinese characters are similar unit to English words in following Zipf's law using short texts (Deng et al., 2014). However, neither studies clarified how well each unit fits. Wang et al. (2005) found that Chinese character frequency obeys Zipf's law in texts written before Qin dynasty but not anymore afterwards. The authors attributed the change to the unification of Chinese characters in Qin dynasty, which leaves little room for the growth of new characters.

Ha et al. (2003) showed the distribution of single characters fall below the expected Zipf's curve in Mandarin news corpora. However, bigram curve fits Zipf's law better than any other n-grams. Moreover, when all n-grams are combined, the data approximately followed Zipf's law for nearly all ranks. Chau et al. (2009) found similar patterns in the distribution of Chinese characters in web searching. They found that bigrams fit the Zipfian distribution better than other n-grams. In addition, the combined n-grams also approximately follow Zipf's law.

In Tibetan, there are super character (i.e. a cluster of consonants and vowels), syllable (a combination of one to seven phonemes), and words (Liu et al., 2014). It was found that syllable and word fit Zipf's law while super character does not, when n-grams from unigrams to 5-grams are combined.

There are also units beyond characters, words and phrases in internet languages. It was also shown that the distribution of hashtags on twitter follows Zipf's law (Melián et al., 2017; Chen et al., 2015). The tags in Chinese blogs also approximately fit Zipf's law (Liu, 2008). In addition, the number of microblog reposts on Sina Weibo obeys Zipf's law (Zhang et al., 2015).

Some non-word symbols also obey Zipf's law. For example, the frequency of emoji used in the discussion of a topic on Chinese microblogging platform follows Zipf's law (Liu et al., 2020). Punctuation in novels written in six Indo-European languages is very similar to words in obeying Zipf's law (Kulig et al., 2017). Williams et al. (2017) further showed that whitespace should also be considered as a word in Zipf's law. Furthermore, both studies showed that when punctuation is added to the analysis, the discrepancy between the power-law and the shifted power-law is resolved.

3 Research method

3.1 Data

The current study used two datasets. The first dataset contains longitudinal danmu comments in a single video and the second dataset includes danmu comments from different videos in 8 categories.

All data come from Bilibili.com, which is the most popular danmu-supported video sharing site in China. The 2020 fourth quarter and fiscal year financial results published by Bilibili Inc. showed that the average monthly active users (MAUs) reached 202 million and the average daily active users (DAUs) rose to 54 million (Bilibili, 2021).

The first dataset was collected by the author and the video (https://www.bilibili.com/video/BV1HJ411L7DP) was selected for three reasons. First, it underwent an abrupt growth. The video was originally uploaded in January 2020, but did not become popular until November 2020. As of March 26, 2021, the video has been watched for more than 30.6 million times and the audience has created more than 90000 danmu comments. Second, the video and its danmu comments have popularized numerous internet buzzwords such as *haoziweizhi* 'mouse tail juice' and *bujiangwude* 'have no martial ethincs'. Finally, the video was relatively recent so the danmu comments can be traced back to the first day the video was published. The Bilibili official API was used to scrape the historical danmu comments between January 5 2020 and March 31 2021. After removing the duplicated items, the author obtained 48459 danmu comments, which is almost half as many as the total danmu comments in the video. Because the Bilibili official API only allows a maximum of 1000 danmu comments for each day and some danmu comments can repeat on different days, it is impossible to get the complete danmu comments.

The second dataset was compiled by the Big Data Lab at University of Science and Technology of China (Lv et al., 2016; Lv et al., 2019). 7.9 million danmu comments were crawled from 4435 videos in 8 categories: anime, movie, dance, music, play, technology, sport, and show. On average, each video provides 1786 danmu comments. Although the authors did not report, it is very likely that the second dataset only collects live danmu comments instead of the complete historical danmu comments as did in the first dataset.

It is worth noting that the two datasets have three major differences: topical homogeneity/heterogeneity, temporal homogeneity/heterogeneity, and size. First, one dataset comes from a single video while the other is extracted from videos of mixed categories. Second, the first dataset contains diachronic danmu comments but the second dataset only includes synchronic danmu comments. Finally, the second dataset is much larger than the first one.

3.2 Hypotheses

The current study postulates two hypotheses regarding the meaningfulness and unit of Zipf's law. The hypotheses will then be tested on the two danmu datasets.

The first hypothesis states that Zipf's law in languages is not a random process. Instead, Zipf's law must be associated with semantics because we use words to express meanings (Manin, 2008; Piantadosi, 2014). This hypothesis is based on three arguments against the random account for Zipf's law.

First, almost all the studies generated random texts with the assumption that each symbol appears with equal probability and thus the frequency of a sequence should decrease monotonically with its length (Manin, 2008). However, this is not the case for natural languages. It was shown that words with three letters are the most frequent in both English and Swedish (Sigurd et al., 2004). Russian data also showed that words with five to ten letters are used most frequently (Manin, 2008). Second, even though random texts may exhibit Zipf's law-like distribution, the distribution in random texts is not identical to real texts. When words are restricted to a certain length, random texts no longer have the Zipfian distribution (Cancho and Solé, 2002). Random texts can also be easily differentiated from natural texts by vocabulary growth (Cohen et al., 1997). Third, the words are used by human beings whose behaviors are far from random. Barabási (2005) argued that humans select tasks according to the priority, rather than acting randomly. It was found that human behaviors are characterized by abrupt bursts and long waiting times between the bursts in email communication. This pattern is different from the regular inter-event time predicted by the Possion distribution, which assumes human activities are random. In addition, acting randomly turns out to be a very difficult task for most people (Iba and Tanaka-Yamawaki, 1996). In a psychological experiment, participants were asked to generate 600 random sequences using digits from 1 to 9 (Schulz et al., 2012). It was shown that even at the eighth sequence, participants' choice of digits can be predicted with an average accuracy rate of 27%, whereas the chance performance is 11%.

The second hypothesis assumes that there is a unit that best fits Zipf's law in linguistic data. The unit must be meaningful and directly observable. According to the first hypothesis, Zipf's law is closely related to meaning. Therefore, the unit in Zipf's law should be carriers of meaningful information. Moreover, it was suggested that direct measurements is more likely to yield Zipfian distribution than derived measurements (Li, 2002).

3.3 Predictions for danmu comments

It is predicted that danmu comments will follow Zipf's law because danmu comments are meaningful and not random. Danmu comments are used to communicate opinions and emotions and may also contain meanings independent of the video content. In addition, He and al. (2017) showed danmu comments in the same video entail a herding effect and multiple-burst phenomena. This pattern is similar to the burst phenomenon in Barabási (2005) which contends that human actions are not random. In addition, danmu comments in a single video will fit Zipf's law better than those in mixed videos because danmu comments in a single video is more topically coherent. Williams et al. (2016) argued that Zipf's law occurs in topically coherent texts. For example, dictionaries, encyclopedias, subjects on questions and answers exhibit poor fit of Zipf's law.

In terms of the units, it is predicted that danmu comments or words are the unit of Zipf's law. There are at least three linguistic units: danmu comments, words and n-grams. Both danmu comments and words are meaningful and directly observable. It is thus expected that the frequency of danmu comments or words will conform to Zipf's law.

4 Results and discussion

4.1 Danmu

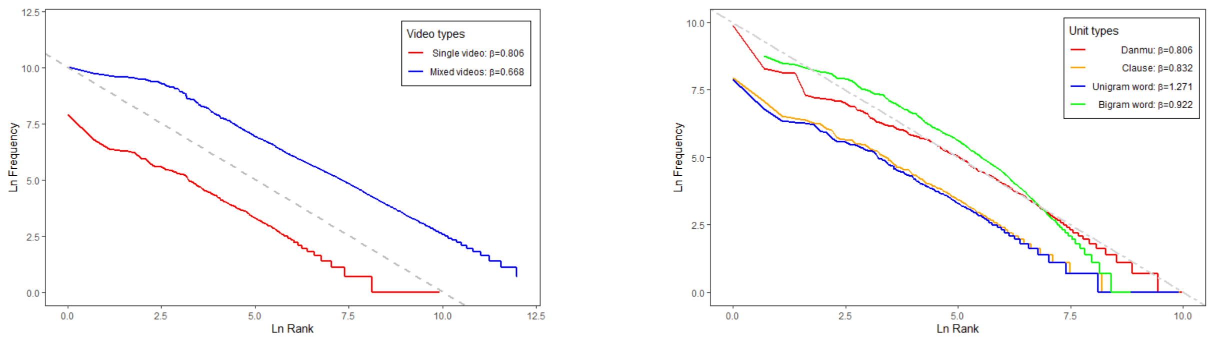
In order to examine whether danmu comments follow Zipf's law, raw danmu comments were used for analysis. No prepossessing such as lowercase conversion was conducted. Table 1 shows the top 10 danmu comments in each dataset.

Single video			Mixed video		
Danmu	Translation	Frequency	Danmu	Translation	Frequency
耗子尾汁	Mouse Tail Juice	2682	卧槽	WTF	22771
很快啊	Very fast	882	完结撒花	The end	17132
哈哈哈	Hahaha	579	233333	A loud laugh	15249
婷婷	Tingting	549	23333	A loud laugh	14118
吭	Onomatopoeia	522	2333333	A loud laugh	14117
全文背诵	Full text recitation	494	哈哈哈哈哈哈	Hahaha	13712
哈哈哈哈	Hahaha	390	哈哈哈哈	Hahaha	12904
不讲武德	No martial ethics	374	23333333	A loud laugh	12824
有bear来	A bear comes	309	哈哈哈	Hahaha	12020
万恶之源	The source of all evil	272	233333333	Hahaha	11671

Table 1: Top 10 danmu comments in two datasets

Note that the two datasets differ in diversity and size. The first dataset came from a single video while the second dataset was taken from more than 4000 videos. In addition, the first dataset has only 46754 items but the second dataset contains 7.9 million items. If danmu comments follow Zipf's law in the same way as random texts, we should expect to see the second dataset fits Zipf's law better because it has higher diversity and larger size.

There are two methods to fit Zipf's law: ordinary least squares (OLS) and maximum likelihood estimation (MLE). It was shown MLE fits better than OLS for both Chinese and English data (Lu et al., 2012). In the current study, the Zipf's exponent β was obtained by fitting the



(a) (b)

Figure 1: Distribution of danmu, clauses, and words in log-log plots. Subfigure (a) shows the distribution of danmu in two datasets. Subfigure (b) shows the distribution of danmu, clause, and words distribution in a single video.

log-transformed data using MLE. Mandelbrot's revision (Equation 2) was used to derive the likelihood function below:

$$L(\theta|r) = \prod_{r=1}^{N} f_r * \frac{1/(r+m)^{\beta}}{\sum_{r=1}^{N} 1/(r+m)^{\beta}} \tag{3}$$

The frequency and rank of danmu comments were log-transformed and plotted in Figure 1a. The exponent in single video danmu is 0.806 and the exponent in mixed video danmu is 0.668. The baseline value for β is 1, as represented by the dashed line.

The results suggest that danmu comments also approximately obey Zipf's law. More importantly, danmu comments in single video fits Zipf's law better. This finding is compatible with Williams et al. (2016), who proposed that Zipf's law is the result of coherent language production and thus topically coherent texts tend to fit Zipf's law better.

So far we have some evidence for the Zipfian distribution of raw danmu comments in the first dataset. In the next three sections, other linguistic units including clauses, words, and n-grams will be analyzed. Due to space constraints, only the results in the first dataset will be reported.

4.2 Clauses

Danmu comments have various lengths. Some are composed of clauses or sentences, such as 没有闪，我笑了"Did not dodge, it was funny" and 这还用问怎么回事？看右眼啊！"Why are you asking what's going on? Just look at the right eye." Thus, clauses can serve as an intermediate stage between danmu comments and words.

The current study uses comma, period, colon, semi-colon, question mark, exclamation mark, and ellipsis in both Chinese and English are as delimiters to cut danmu into clauses. The frequency of the clauses is fitted to Mandelbrot's revision (Equation 2) using MLE. The exponent β of clauses is 0.832, slightly closer to 1 than the raw danmu.

4.3 Words

Although methods for Chinese word segmentation are mature, the current study used unigrams and bigrams of Chinese characters as a proxy to Chinese words.

This practice is adopted for two reasons. First, word-segmentation can be difficult to apply to mixed language. Code-mixing and language variants are very common in Internet language. In Table 1, "有bear来" is an example of code mixing and language variants can be found in "哈哈哈", "哈哈哈哈", "哈哈哈哈哈" of different lengths. Second, using characters allows the current study to be comparable to previous studies such as Chau et al. (2009), who investigated character usage in Chinese web searching.

Before analyzing word frequency, numbers and punctuation were removed from the danmu comments. Strings in Chinese, Japaneses, and Korean were segmented by character or syllable. String in Indo-European languages were segmented by whitespaces. Contractions such as "'s" and "n't" were restored to the original words as "is" and "not".

The unigrams and bigrams are a mixture of different languages, including Chinese, Japanese, Korean, English, French, German, etc. Linear regression was used to calculate Zipf's exponent for each linguistic unit. As shown in Figure 1b, the exponent β in danmu, clauses, unigrams, and bigrams are 0.806, 0.832 1.271, and 0.922 respectively. Evidently, bigrams best fit Zipf's law. This finding is also consistent with Chau et al. (2009), Ha et al. (2003), and Williams et al. (2015). The studies found that bigrams follow Zipf's law better than unigrams in Chinese and phrases fit Zipf's law better than words in English. Furthermore, the exponent in bigrams is less deviated from 1 than that in danmu comments, though both are close to 1. This may suggest that there are different units for Zipf's law in Chinese internet language, but bigrams or words are still the most basic linguistic unit.

The implications of the findings are twofold. It suggests that multiple linguistic units in the same data can fit Zipf's law, similar to the distribution of lemmas and words in Indo-European languages (Kwapień and Drożdż, 2012; Corral et al., 2015). The findings also filled a gap in previous studies on Zipf's law in internet language. The internet language potentially involves different linguistic units, however, those different units were not directly compared in previous studies. For example, character usage in web searching was looked into without comparing it to the query terms (Chau et al., 2009). Frequencies of hashtags and blog tags were also examined, but no comparison has been made with regard to characters or words that constitute the hashtags or tags (Liu, 2008; Chen et al., 2015; Melián et al., 2017).

4.4 N-grams

Another issue that needs to be addressed is n-grams. According to the hypotheses of unit in Zipf's law proposed in the current study, the unit has to be meaningful and directly observable. However, this is not always the case for n-grams. N-grams are derived data and sometimes do not make sense. Consequently, they should not be considered as the unit for Zipf's law. However, several studies reported that the combined n-grams fit Zipf's law for almost all ranks regardless of the languages and the unit chosen for creating n-grams (Ha et al., 2003; Ha et al., 2009; Chau et al., 2009).

Ha et al. (2009) provided an explanation for the behavior of combined n-grams with randomly generated bits. The paper claimed that the extension of Zipf's law to n-grams may arise from pure probabilities and called for an re-examination of the theoretical motivation of Zipf's law.

To attest the argument, the current study compared separated n-grams and combined n-grams of different symbols, including words, numbers, and punctuation. The n-grams for these symbols were extracted from each danmu comment and then merged together.

The log-log plots for different symbols are shown in Figure 2 and the Zipf's exponents were presented in Table 2. As can be seen in the table, the combined n-grams have smaller β values than unigrams in all three types of symbols. This characteristic is the same as reported in previous studies.

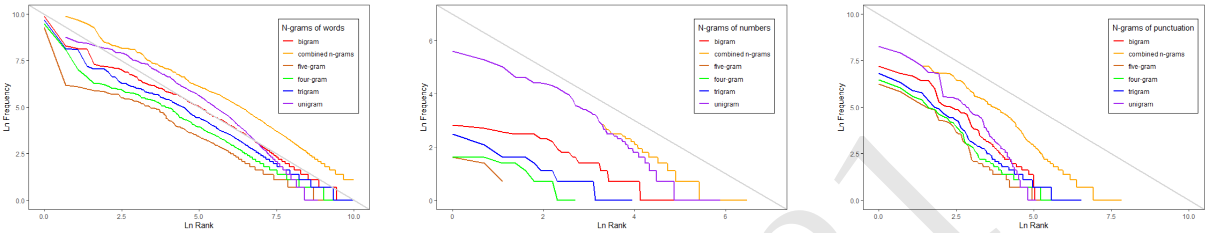
However, after taking a closer look, the advantage of combined n-grams is not so ubiquitous. In the column of punctuation n-grams, the combined n-grams are only better than the unigrams and bigrams. The β value of combined n-grams of punctuation is also much higher than that of words and number.

Punctuation usually includes one symbol such as "?" and some may contain two or three symbols such as "[ ]" and "!!!". It is hard to interpret what 4-grams and 5-grams of punctuation mean, yet their exponents are all closer to -1 than the combined n-grams. A possible explanation is that the combined n-grams may be a statistical smoothing. Because punctuation data have few 4-grams and 5-grams, the smoothing does not have a strong effect. On the contrary, the combined n-grams for words are much for effective.

In summary, combined n-grams are the best fit for Zipf's law for both words and numbers. However, the somewhat meaningless punctuation 4-grams and 5-grams fit Zipf's law better than the combined n-grams.

	Words	Numbers	Punctuation
Zipf's exponent of 1-gram	1.271	1.465	2.305
Zipf's exponent of 2-gram	0.922	0.764	1.708
Zipf's exponent of 3-gram	0.842	0.692	1.255
Zipfs' exponent of 4-gram	0.775	3.282	1.250
Zipf's exponent of 5-gram	0.687	4.997	1.306
Zipf's exponent of combined n-grams	0.956	1.155	1.444

Table 2: Zipf's exponent for n-grams of different symbols



(a) (b) (c)

Figure 2: N-grams of different symbols in log-log plots. Subfigure (a), (b), (c) shows N-grams of words, numbers, and punctuation respectively.

5 Conclusion

The current study explored Zipf's law with novel datasets from danmu comments, which is an emerging type of online video comments and internet language.

Specifically, the study has three findings. First, danmu comments also follow Zipf's law and danmu comments from topically homogeneous video fit Zipf's law better. The findings suggest that Zipf's law may be driven by semantic motivations. Second, danmu comments, clauses, and character unigrams and character bigrams in danmu comments all follow Zipf's law. There seems to be a continuum of fitness in the order of bigrams, clauses, and danmu comment. The results indicate that multiple units for Zipf's law may exist even in the same data. Finally, different from previous studies, combined n-grams do not always have the best fit for Zipf's law. It is argued that combined n-grams may be considered as a statistical smoothing rather than a manifestation of Zipf's law.

There are two limitations of the datasets. First, danmu comments are cumulative and may change over time. Therefore, a temporal analysis of danmu comments is necessary in the future. In addition, the danmu comments from a single video seems to fit Zipf's law. However, the dataset is relatively small and larger data from more videos are needed to validate the Zipfian distribution in danmu comments.

Internet has greatly changed the way we use languages and multilingualism is becoming increasingly popular. Studies on Zipf's law in internet language will help us understand the language phenomena better.

Acknowledgements

The author is grateful to the three anonymous reviewers for providing helpful feedback on this paper. The author also wants to thank the BigData Lab at University of Science and Technology of China for providing the danmu data.

References

Qingchun Bai, Qinmin Vivian Hu, Linhui Ge, and Liang He. 2019. Stories that big danmaku data can tell as a new media. *IEEE Access*, 7:53509–53519.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science*, 286(5439):509–512.

Albert-Laszlo Barabasi. 2005. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211.

Bilibili. 2021. 2020 fourth quarter and fiscal year financial results. Available at https://ir.bilibili.com/static-files/09f30d5d-5de5-4338-b767-921ce1a07a47 (2021/03/26).

Ramon Ferrer i Cancho and Ricard V Solé. 2002. Zipf's law and random texts. *Advances in Complex Systems*, 5(01):1–6.

Ramon Ferrer i Cancho and Ricard V Solé. 2003. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791.

Michael Chau, Yan Lu, Xiao Fang, and Christopher C Yang. 2009. Characteristics of character usage in chinese web searching. *Information processing & management*, 45(1):115–130.

Yue Chen, Qin Gao, and Pei-Luen Patrick Rau. 2015. Understanding gratifications of watching danmaku videos–videos with overlaid comments. In *International Conference on Cross-Cultural Design*, pages 153–163. Springer.

Ye-Sho Chen. 1991. Zipf's law in natural languages, programming languages, and command languages: the simon-yule approach. *International journal of systems science*, 22(11):2299–2312.

Albert Cohen, Rosario Nunzio Mantegna, and Shlomo Havlin. 1997. Numerical analysis of word frequencies in artificial and natural language texts. *Fractals*, 5(01):95–104.

Álvaro Corral, Gemma Boleda, and Ramon Ferrer-i Cancho. 2015. Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PloS one*, 10(7):e0129031.

Carlos R Cunha, Azer Bestavros, and Mark E Crovella. 1995. Characteristics of www client-based traces. Technical report, Boston University Computer Science Department.

Weibing Deng, Armen E Allahverdyan, Bo Li, and Qiuping A Wang. 2014. Rank-frequency relation for chinese characters. *The European Physical Journal B*, 87(2):1–20.

Robert MW Dixon and Alexandra Y Aikhenvald. 2002. Word: a typological framework. In Robert MW Dixon and Alexandra Y Aikhenvald, editors, *Word: A cross-linguistic typology*, pages 1–41. Cambridge University Press.

Xavier Gabaix. 1999. Zipf's law for cities: an explanation. *The Quarterly journal of economics*, 114(3):739–767.

Yi Guan, Xiaolong Wang, and Kai Zhang. 1999. Xiandai hanyu jisuan yuyan moxing zhong yuyan danwei de pindu—pinji guanxi [the frequency-rank relation of language units in modern chinese computational language model]. *Zhongwen Xinxi Xuebao [Journal of Chinese Information Processing]*, 13(2):9–16.

Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming, and F. J. Smith. 2003. Extension of zipf's law to word and character n-grams for english and chinese. *Computational Linguistics and Chinese Language Processing*, 8:77–102.

Le Quan Ha, Philip Hanna, Ji Ming, and FJ Smith. 2009. Extending zipf's law to n-grams for large corpora. *Artificial Intelligence Review*, 32(1):101–113.

Ming He, Yong Ge, Enhong Chen, Qi Liu, and Xuesong Wang. 2017. Exploring the emerging type of comment for online videos: Danmu. *ACM Transactions on the Web (TWEB)*, 12(1):1–33.

Yukito Iba and Mieko Tanaka-Yamawaki. 1996. A statistical analysis of human random number generators. In *Proceedings of the 4th International Conference on Soft Computing (IIZUKA'96), Iizuka, Fukuoka, Japan, Sep*, pages 467–472. Citeseer.

Andrzej Kulig, Jarosław Kwapień, Tomasz Stanisz, and Stanisław Drożdż. 2017. In narrative texts punctuation marks obey the same statistics as words. *Information Sciences*, 375:98–113.

Jarosław Kwapień and Stanisław Drożdż. 2012. Physical approach to complex systems. *Physics Reports*, 515(3-4):115–226.

Wentian Li. 1992. Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on information theory*, 38(6):1842–1845.

Wentian Li. 2002. Zipf's law everywhere. *Glottometrics*, 5:14–21.

Ruxin Li. 2018. Shipin danmu de yuyanxue yanjiu [a linguistic analysis of danmu comments]. Master's thesis, Shaanxi Normal University.

Huidan Liu, Minghua Nuo, and Jian Wu. 2014. Zipf's law and statistical data on modern tibetan. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 322–333.

Fei Liu, Hao Wang, and Xiaoke Xu. 2020. Shejiaomeiti zhong biaoqingfuhao de shiyong xingwei ji chengyin fenxi [usage behavior and cause of emoji in social media]. *Fuza Xitong Yu Fuzaxing Kexue [Complex Systems and Complexity Science]*, 17(3):70–77.

Zhiyuan Liu. 2008. Chinese language networks research and applications. Master's thesis, Tsinghua University.

Gaofei Lu, Pu Han, and Si Shen. 2012. Comparative empirical study on zipf's law with two fitting methods. *Library and information service*, 56(24):71–76.

Guangyi Lv, Tong Xu, Enhong Chen, Qi Feng Liu, and Yi Zheng. 2016. Reading the videos: Temporal labeling for crowdsourced time-sync videos based on semantic embedding. *AAAI*, pages 3000–3006.

Guangyi Lv, Kun Zhang, Le Wu, Enhong Chen, Tong Xu, Qi Liu, and Weidong He. 2019. Understanding the users and videos by mining a novel danmu dataset. *IEEE Transactions on Big Data*.

Bill Z Manaris, Luca Pellicoro, George Pothering, and Harland Hodges. 2006. Investigating esperanto's statistical proportions relative to other languages using neural networks and zipf's law. In *Artificial Intelligence and Applications*, pages 102–108.

Benoit B Mandelbrot and Benoit B Mandelbrot. 1982. *The fractal geometry of nature*, volume 1. WH freeman New York.

Benoit Mandelbrot. 1953. An informational theory of the statistical structure of language. *Communication theory*, 84:486–502.

Dmitrii Y Manin. 2008. Zipf's law and avoidance of excessive synonymy. *Cognitive Science*, 32(7):1075–1098.

Ali Mehri and Maryam Jamaati. 2017. Variation of zipf's exponent in one hundred live languages: A study of the holy bible translations. *Physics Letters A*, 381(31):2470–2477.

José Alberto Pérez Melián, J Alberto Conejero, and Cesar Ferri Ramirez. 2017. Zipf's and benford's laws in twitter hashtags. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 84–93.

Seio Nakajima. 2019. The sociability of millennials in cyberspace: A comparative analysis of barrage subtitling in nico nico douga and bilibili. In Vanessa Frangville and Gwennaël Gaffric, editors, *China's Youth Cultures and Collective Spaces: Creativity, Sociality, Identity and Resistance*, pages 98–115. Routledge.

Mark EJ Newman. 2005. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351.

Steven T Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130.

Sidney Redner. 1998. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2):131–134.

Yukie Sano, Hideki Takayasu, and Misako Takayasu. 2012. Zipf's law and heaps' law can predict the size of potential words. *Progress of Theoretical Physics Supplement*, 194:202–209.

Marc-André Schulz, Barbara Schmalbach, Peter Brugger, and Karsten Witt. 2012. Analysing humanly generated random number sequences: a pattern-based approach. *PloS one*, 7(7):e41531.

Bengt Sigurd, Mats Eeg-Olofsson, and Joost Van Weijer. 2004. Word length, sentence length and frequency–zipf revisited. *Studia linguistica*, 58(1):37–52.

Herbert A Simon. 1955. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440.

Dahui Wang, Menghui Li, and Zengru Di. 2005. True reason for zipf's law in language. *Physica A: Statistical Mechanics and its Applications*, 358(2-4):545–550.

Jake Ryland Williams and Giovanni C. Santia. 2017. Is space a word, too? *arXiv preprint arXiv:1710.07729*.

Jake Ryland Williams, Paul R Lessard, Suma Desu, Eric M Clark, James P Bagrow, Christopher M Danforth, and Peter Sheridan Dodds. 2015. Zipf's law holds for phrases, not words. *Scientific reports*, 5(1):1–7.

Jake Ryland Williams, James P Bagrow, Andrew J Reagan, Sharon E Alajajian, Christopher M Danforth, and Peter Sheridan Dodds. 2016. Zipf's law is a consequence of coherent language production. *arXiv preprint arXiv:1601.07969*.

Qunfang Wu, Yisi Sang, and Yun Huang. 2019. Danmaku: A new paradigm of social interaction via online videos. *ACM Transactions on Social Computing*, 2(2):1–24.

Yaxing Yao, Jennifer Bort, and Yun Huang. 2017. Understanding danmaku's potential in online video learning. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, pages 3034–3040.

Ning Zhang, Shuqing Zhang, Hong Chen, and Yang Luo. 2015. Xinlang weibo zhuanfashu de milü fenbu xianxiang [the power-law distribution in the number of reposts in sina microblog]. *Jisuanji Shidai [Computer era]*, 3:33–35.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley Press, Cambridge.

Peter Zörnig. 2015. Zipf's law for randomly generated frequencies: explicit tests for the goodness-of-fit. *Journal of Statistical Computation and Simulation*, 85(11):2202–2213.