How much is said in a microblog? A multilingual inquiry based on Weibo and Twitter

Han-Teng Liao
Oxford Internet Institute
University of Oxford
hanteng@gmail.com

King-wa Fu
Journalism and Media Studies Centre
University of Hong Kong
kwfu@hku.hk

Scott A. Hale
Oxford Internet Institute
University of Oxford
scott.hale@oii.ox.ac.uk

ABSTRACT

This paper presents a multilingual study on, per single post of microblog text, (a) how much can be said, (b) how much is written in terms of characters and bytes, and (c) how much is said in terms of information content in posts by different organizations in different languages. Focusing on three different languages (English, Chinese, and Japanese), this research analyses Weibo and Twitter accounts of major embassies and news agencies. We first establish our criterion for quantifying "how much can be said" in a digital text based on the openly available Universal Declaration of Human Rights and the translated subtitles from TED talks. These parallel corpora allow us to determine the number of characters and bits needed to represent the same content in different languages and character encodings. We then derive the amount of information that is actually contained in microblog posts authored by selected accounts on Weibo and Twitter. Our results confirm that languages with larger character sets such as Chinese and Japanese contain more information per character than English, but the actual information content contained within a microblog text varies depending on both the type of organization and the language of the post. We conclude with a discussion on the design implications of microblog text limits for different languages.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation (e.g., HCI)**]: Group and Organization Interfaces—*Web-based interaction*; H.5.4 [**Information Interfaces and Presentation (e.g., HCI)**]: Hypertext/Hypermedia

General Terms

Human Factors; Measurement

Keywords

Microblogs; Language; Design; Social Networking

1. ARBITRARY CHARACTER LIMITS?

Microblogging platforms are distinguished from traditional blogging platforms by having a limit on the length of posts. Such length limitations are reported to lower the time and thought required to create new content and allow for faster and more timely exchange of information [6, 16]. Most studies considering the effects of length limitations on the user experience of microblogging have examined English-language content, but there is ample reason to believe that universal length limitation have different effects on people writing content in different languages. It has been reported by major media such as BBC [13] and the Atlantic [29] that one can express much more content in languages other than English within a given character limit, most notably the 140-character limit of Twitter and the 140-byte limit of Short Message Service (SMS, or text messages). Chinese [2] and Japanese [33] are often cited as examples of "more expressive" languages within such space limits.

However, a definitive answer using systematic and quantifiable methods is yet to be provided to the question of how much more expressive a given language is within such length limitations. Neubig and Duh [26] provide the only academic work on this subject and use an information-theoretic approach. They find that Chinese and Japanese are the most expressive languages per character, but do not use parallel corpora (i.e., the same information in multiple languages) in their work. Apart from academic scholarship, the issue of language expressiveness has received much attention in the popular press with the BBC reporting that 140 Chinese characters amounts to 70 to 80 English words [13]. Bloggers have also weighed in on the debate with one blogger writing that 140 Chinese characters could contain five times more content than the same number of English characters [30] and another blogger claiming that "140 Chinese characters is more like 500 characters on Twitter.com" [5]. By machine translating foreign-language content from Twitter from "a few users," IT Consultant Ben Summers reported on his blog that Japanese tweets could contain information that would take up to 260 English characters to express [33]. These differing estimates and the lack of academic scholarship on this topic motivate our paper.

The imposition of length constraints has profound effects on how platforms are used and hence the user experience of these platforms. The effects of length constraints on users are not new or unique to microblogging platforms. The user experience of SMS has received much attention, and scholars have found that the limitations resulted in specific language practices including specific abbreviations (e.g., b4,

2day) as well as more contractions in comparison to instant messaging [21]. It is difficult to disentangle length and input device limitations when comparing instant messaging and SMS [21], but one study found SMS messages sent via Internet-connected PCs were longer than those sent directly from mobile (feature) phones [10]. Finally, Grinter et al. [10] found that the length limitations of SMS allowed teenagers to forego conversational conventions and reduce the overall time spent on interactions.

In this paper, we attempt to answer the general questions of how space/length restrictions affect (a) how much can be said, (b) how much space is used, and therefore (c) how much is actually said within imposed space limitations and how these factors differ across languages. These basic questions are central to any cross-lingual assessment of the impact of length limitations on microblog posts and hence the user experience of people on such services and have important ramifications for designers of microblogging platforms as well as people using such platforms.

1.1 Reasons for length limitations

Early character limits were often imposed as a result of technical limitations. For example, the character limit of text messages (SMS) was the result of an engineering choice to send SMS messages via the existing signaling channels of the GSM system (but with a lower priority than other signaling messages) [14]. After including header information, the existing channel left 140 octets (bytes) available for the text of an SMS message. Using 7-bit character encoding, 160 Latin characters were made to fit in this space [1]. Hillebrand, an early pioneer of SMS, reasoned that 160 characters would be sufficient through an examination of the number of characters (including spaces and punctuation marks) for random short paragraphs [9, 34]. Although we cannot be certain which language he used for this exercise, it is likely to have been English or German. If another language (particularly a language with a non-Roman alphabet such as Japanese or Chinese) had been used, Hillebrand may have reached a different conclusion about the feasibility of using the existing signaling channels of the GSM system and ultimately developed SMS in a different manner with a different character limit.

The continuation of character limits on modern microblogging platforms is less a case of technical limitations and more often a design choice aiming to cultivate a certain user experience. Best practices published by Twitter in 2013 mentioned the 140-character limit and stated that "creativity loves constraints and simplicity is at our core." [36]. Character limitations may have very different effects in different languages as cross-platform and cross-lingual analysis of microblogs suggests language is an important factor for researchers to consider [e.g., 7, 11, 12, 15, 19].

When Twitter launched, it set a limit of 140 characters for posts so that a Twitter post would fit within a single SMS message and leave 20 characters for usernames and other commands [9, 23]. It should be noted, however, that only for 140 English characters would this have ever been the case as SMS itself has a hard limit of 140-bytes within a single message as previously stated. Thus, a single SMS message can only accommodate 70 Chinese or Japanese characters. The limits imposed on Twitter, however, are character limits and not byte limits [37], and thus, it is possible to send a tweet with 140 Chinese characters even though such a tweet would never have fit within one SMS message. Now that most Twitter traffic comes from native apps on smartphones, which have no technical limitation on the length of messages, the continued imposition of the 140-character limit is due less to technical reasons and more to user-experience design choices.

When Sina Weibo launched in mainland China it also imposed a limit on the length of posts. This limit is often reported to also be 140 characters [e.g., 5], but the data we collected in this study revealed several posts with more than 140 characters. Our experimentation with the Sina Weibo interface indicates that Sina Weibo likely imposes a byte limit and not a character limit. We found that posts on Sina Weibo can be up 140 Chinese characters or 280 English/Latin characters.[1] In other words, Sina Weibo imposes a varying limit on the number characters a message may contain depending on the number of bytes required to store each character. In practice, this means that messages in English on the platform could be twice as long as messages in Chinese, with most other languages having maximum lengths somewhere between these endpoints depending on the frequency of accented and other special characters in the language.

Apart from the question of the actual limits imposed on different communication platforms, marketers and other users of microblogging platforms have asked what the "ideal" message length is for driving strong engagement, with different numbers reported for different platforms including Twitter, Facebook and Google Plus [17]. For example, it is reported and recommended by various industry research organizations that the ideal length of a tweet is either 100 or 71–100 characters [17, 35]. It is also reported that the ideal Facebook post be around 40 characters while the Google Plus posts should be around 60 characters [35], despite these platforms allowing much longer posts in practice [3]. However, despite the marketing and industry interest in the factor of microblog post length for engagement optimization, there is little research on length across different languages and/or different platforms. In other words, we need to fill the gap in both research and design with cross-cultural considerations [34] regarding both the limits and actual practices regarding the length of posts.

1.2 A parallel corpora approach

Corpus linguists have partially addressed the question on how much can be said in different languages using parallel corpora. Before Twitter came into existence, the Director of Linguistic Data Consortium (a major linguistic research data keeper) conducted a well-designed comparison of Chinese and English using the LDC parallel Chinese/English corpora, producing ratio results ranging from 1.96 to 2.27 for texts and 1.19 to 1.24 for compressed (gzipped) files [20]. The underlying idea behind Liberman's research design was straightforward. By comparing the actual storage space of texts of the same content in different languages, one can see how much "space" is required to store/convey the same content. This idea is expressed by the equation below:

$$S_{lang_A} \times IC_{lang_A} = S_{lang_B} \times IC_{lang_B} \quad (1)$$

[1]This suggests Sina Weibo imposes a byte limit with a character encoding such as GBK that uses one byte to store most English/Latin characters and two bytes to store other characters including most Chinese characters.

S_{lang} denotes the "space" units required to store the content in language *lang*. IC_{lang} denotes the amount of "information content" per space unit in language *lang*. The equation holds only for parallel corpora, which by definition contain the same content in each language (i.e., where $lang_A$ and $lang_B$ contain the same information written in different languages).

While Liberman's research is systematic and addresses the computational questions of information storage, it is of limited application to modern social media data for two reasons. First, the amount of storage space required to store information in different languages depends on the character encoding used. Liberman's work used the Chinese national standard GB-2312 and not the international Unicode encodings, which have become more common for multilingual websites and applications. Second, the data Liberman used was formal and legalistic in nature. The ratios between such formal text may not apply to the more conversational and informal text commonly found on modern social media platforms.

2. METHODS AND DATA SELECTION

We propose a systematic step-wise approach that can be extended to cover more platforms, languages, and types of microblogs, with the aim to answer our three main research questions. We choose to focus on news and diplomacy organization accounts because we expect the length of messages would be integral to the overall communication strategies, including in potential cross-lingual or cross-cultural scenarios. The data selection in this paper is limited to two platforms, three languages, and the most recent posts of 54 microblogs. However, as the first multiplatform and multilingual study its contributions are important and lay the foundation for future work with additional account types, languages, and platforms.

The following sections describe the three steps of the research process, each of which answers one of our research questions concerning (a) how much can be said, (b) how much is typed and posted, and (c) how much is said in different languages by different organizations. First, we calculate the cross-lingual ratios of information content.

2.1 Calculating the cross-lingual ratios of information content based on UDHR and TED talks

Following the approach of Liberman's corpus linguistic study [20], we propose a generic research design to measure the ratios of "information content" in different languages.

The ratio of information content per space for language B to language A can be derived from Equation 1 to produce the ratio given in Equation 2. This is the inverse ratio of "space" required to store the same content when a parallel corpus is used. In other words, if more space is required to store the same content for language B than language A, the ratio value will be less than one, indicating the information content per space unit of language B is smaller than that of language A.

$$ratio(lang_B, lang_A) = \frac{IC_{lang_B}}{IC_{lang_A}} = \frac{S_{lang_A}}{S_{lang_B}} \qquad (2)$$

The potential of using Web content as a parallel corpus for research has been proposed [28] and executed [18]. We use a parallel corpus formed from human-translated user-generated content for several reasons. First, the content is open and freely available providing for easier replication in comparison to conventional parallel linguistic corpora, which often require license fees to use. Second, the human translation of the content usually provides better quality text than corpora formed with machine translation of open content. Third, most user-generated content is generally more up-to-date and contemporary than that of conventional corpora and is closer in style to the text used on microblogging platforms and thereby provides a more suitable basis for the research on microblogging.

We analyze two corpora to understand how corpus selection influences our results. The first corpus we use is the UDHR in Unicode Project, which provides translations of the Universal Declaration of Human Rights (UDHR) to demonstrate the use of Unicode for multilingual environments. Because of the normative, universal and semilegal status of UDHR, the Unicode translation project is the "most translated text" [38]. Thus, the project can provide parallel corpora that cover the most languages in the world for comparison.

Our second parallel corpus is formed from the TED Open Translation Project. This project is led by a well-known, Internet-friendly organization and uses professional human translation service to kick-start the crowd-sourced translation of video subtitles. The translated video subtitles form a multilingual corpora that is comparatively closer in style to informal, online communication. Of course speeches on TED.com are still different from online expressions such as microblog posts, but they are closer to online expressions when compared to legal or governmental data that is commonly used in corpus linguistics because of its institutional availability.

To calculate the ratios of information content across languages, we downloaded the texts from the UDHR in Unicode Project and subtitles for all of the 1,847 TED talks available from TED.com (a complete sample of all videos available as of March 16, 2015). We have downloaded the available transcripts of four language versions: English, Japanese, simplified Chinese, and traditional Chinese. Our full analysis of the TED talks include 1,522 videos from all the videos available because 209 videos did not have transcripts available in all four languages and a further 116 videos had extremely short lengths (these were mostly performance videos).

The texts were parsed and the number of characters in each video in each language was determined. For the UDHR datasets, we use each paragraph as a unit to calculate the respective ratios across languages and then produce basic descriptive statistics. For the TED datasets, we use each speech (i.e., all the subtitles for one video) as a unit.

2.2 Measuring the text lengths of posts from selected accounts from Twitter and Weibo

To measure the text lengths of microblog posts, we examine the posts by 54 news and embassy organizations.

For Twitter, we selected 36 accounts that post either in English, Japanese, or Chinese. The most recent 200 tweets for each account were collected through the Twitter REST API on 22 January 2015. For Weibo, we collected samples from a data intermediary called Weiboscope [7, 8] that provided uncensored and randomly sampled datasets from Weibo. The data was first collected by making SQL requests

to the database for messages posted between 1 January and 22 January 2015. To ensure the number of posts per account was large enough, we only considered accounts having more than 50 posts. Two Weibo accounts that are owned by the World Bank and the Economist (with respective screen names "世界银行" and "经济学人集团") were thus excluded, leaving us with 18 accounts for analysis. The first column of Table 1 lists the screen names of the Twitter and Weibo accounts analyzed.

With the help of human readers and our own language identification algorithms, we found that almost all accounts used one language. We coded each account accordingly with their respective language codes. We found that two Twitter accounts posted content in both English and Japanese ("UKinJapan" and "usembassytokyo"). We collected 100 posts in each language for each of these accounts, and analyze the content in each language separately. Each of these accounts is marked with an asterisk (*) in Table 1.

We coded each organization account as either *embassy* or *news* (the *type* column of Table 1). A small amount of gray area exists between these two categories. For example, news organizations such as China's People's Daily ("people_cn") may also be media organs of state governments. Other organizations, such as the UN and World Bank, are not strictly embassies. Nevertheless, these international organizations have their political significance in diplomacy and in some ways their use of microblogs will be similar to the use of microblogs by embassies. At the very least, these embassy and embassy-like organizations provide a contrast in type to news organizations. The selected accounts in Table 1 thus contain a mixture of microblogs written in different languages and belonging to different types of organizations. Although the selection of accounts and posts is not random, the selected accounts cover major news and embassy organizations across three major languages on two major microblogging platforms.

To measure the text lengths, we first remove all hyperlinks in the microblog posts and calculate the number of Unicode characters per post.[2] Based on this measurement we derive the average length in characters (i.e., the mean value of characters per post) for each account (we ignore the differences between single-byte and multi-byte characters as Twitter does in enforcing its character limit [37]). With the average length values across different platforms, languages, and types of organizations, we should find out whether and how such character lengths vary, thereby answering the question of how much is typed and posted.

2.3 Estimating the relative information content in a microblog post for cross-lingual comparison

[2]We remove URLs from posts before making our length comparison. Hong et al. [15] found that tweets in different languages included URLs at different frequencies, but within our restricted set of embassies and news organizations, we find that URLs are included within tweets at a similar rate between the three languages we analyze. The majority of tweets we analyze in each language contained exactly one URL (67% of the Chinese tweets, 69% of the English tweets, and 72% of the Japanese tweets that we analyze had exactly one URL). The analysis presented here strips URLs from all tweets and compares the results, but repeating the analysis with only tweets that contained exactly one URL yields nearly identical findings.

Twitter

Screen name	Language	Type
ft	English	news
ftchina	English	news
KyodoNewsENG	English	news
wsj	English	news
xinhuanetnews	English	news
47news	Japanese	news
asahi	Japanese	news
bbcjapan	Japanese	news
mainichiJapaneseews	Japanese	news
nikkei	Japanese	news
peopledailyjp	Japanese	news
sankei_news	Japanese	news
WSJJapan	Japanese	news
asahi_shinsen	Simplified Chinese	news
bbcchinese	Simplified Chinese	news
china_kyodonews	Simplified Chinese	news
chinesewsj	Simplified Chinese	news
djy_cn	Simplified Chinese	news
dw_chinese	Simplified Chinese	news
people_cn	Simplified Chinese	news
voachina	Simplified Chinese	news
UKinJapan*	English	embassy
UN	English	embassy
usembassytokyo*	English	embassy
worldbank	English	embassy
ChnEmbassy_jp	Japanese	embassy
Embassy_ItalyJP	Japanese	embassy
IcelandEmbTokyo	Japanese	embassy
IsraelinJapan	Japanese	embassy
koreanemb_japan	Japanese	embassy
NLinJapan	Japanese	embassy
UKinJapan*	Japanese	embassy
UKRinJPN	Japanese	embassy
usembassytokyo*	Japanese	embassy
worldbanktokyo	Japanese	embassy
france_in_china	Simplified Chinese	embassy
UNRadioChinese	Simplified Chinese	embassy
usa_china_talk	Simplified Chinese	embassy

Weibo

Screen name	Language	Type
人民网	Simplified Chinese	news
日本共同社	Simplified Chinese	news
韩国中央日报	Simplified Chinese	news
半岛电视台AlJazeera	Simplified Chinese	news
ETtoday新聞雲	Simplified Chinese	news
FT中文网	Simplified Chinese	news
华尔街日报中文网	Simplified Chinese	news
福布斯中文网	Simplified Chinese	news
路透中文网Reuters	Simplified Chinese	news
联合国	Simplified Chinese	embassy
美国驻华大使馆	Simplified Chinese	embassy
加拿大大使馆官方微博	Simplified Chinese	embassy
韩国驻华大使馆	Simplified Chinese	embassy
以色列驻华使馆	Simplified Chinese	embassy
俄罗斯驻华大使馆	Simplified Chinese	embassy
韩国驻华大使馆	Simplified Chinese	embassy
以色列驻华使馆	Simplified Chinese	embassy
俄罗斯驻华大使馆	Simplified Chinese	embassy

Table 1: Twitter and Weibo accounts analyzed in this study

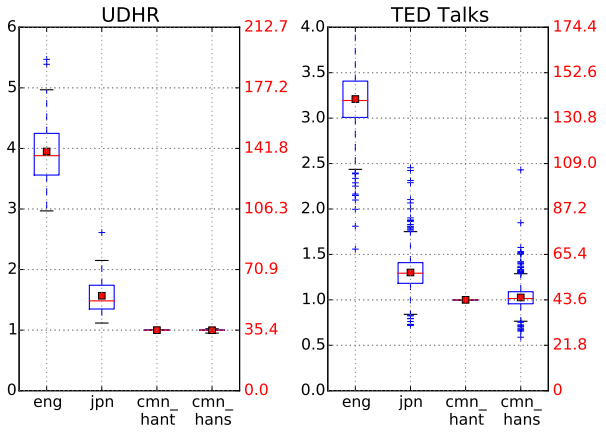

Figure 1: Relative ratio of characters in English (eng), Japanese (jpn), and Chinese using simplified characters (cmn_hans) required to express the same content compared to Chinese using traditional characters (cmn_hant) as the baseline.

After we calculate the ratios of information content and the lengths, we can then derive an estimation of relative information content (RIC) as given in Equation 3.

$$RIC_{lang_A} = \frac{S_{lang_B}}{ratio(lang_B, lang_A)} = S'_{lang_A} \quad (3)$$

The main purpose of the Equation 3 is to allow for cross-lingual comparisons on an equal basis. Using language A as the baseline, the relative information content can be derived by dividing the value of the number of space units (S_{lang_B}) used when the content is expressed in language B with the ratio of information content ($ratio(lang_B, lang_A)$). The formula produces the equivalent space that would be required to expressed the same content in language A (S'_{lang_A}). We present results where space is measured either as characters or as UTF-8 bytes.

3. FINDINGS

The findings are presented in three subsections, each answering one of our three main research questions. The findings of the first subsection confirm that Chinese and Japanese contain more information per character. The findings of the second subsection show a mixed picture on the length of microblog posts by different organizations in different languages. Combining the results from the first two subsections, the third and final subsection presents the derived measurement of the relative information content expressed in microblog posts by different organizations in different languages.

3.1 How much can be said: UDHR and TED talks

Figure 1 shows the outcomes of the research on the UDHR texts and the TED talks respectively using box plots. Both sets of box plots show that indeed, per Unicode character, the Chinese language can express the same idea with fewer characters. Using Chinese written with traditional characters as the baseline, the set of box plots to the left shows for the same paragraph content of UDHR that it takes on average nearly four times as many characters in English to express the same content (the mean value of 3.95 is shown by the red rectangular box in the *eng* column). Similarly, it takes only about 1.6 as many characters to express the same content in Japanese. There is no significant difference in the number of traditional or simplified characters needed to write the same content in Chinese. If the type of language used in the UDHR was typical of the language used in microblog posts, these findings would indicate that a tweet (or a Sina Weibo post) of 140 Chinese characters could convey nearly four times as much information as an English message of the same character length and 1.6 times as much information as a Japanese message of the same character length.

In order to understand the impact of the specific parallel corpus used on these calculations, we repeat the same analysis using the speech content of TED talks and show the results in the right set of box plots in the same figure (Figure 1). This analysis yields slightly lower numbers, showing that it takes on average about 3.2 times the number the characters to express the same content in English in comparison to Chinese. Similarly, it takes about 1.3 times as many characters to express the same content in Japanese in comparison to Chinese. Once again, there is no significant difference between the number of simplified or traditional characters needed to express the same content in Chinese.

Relating these findings to microblogs, we can calculate the equivalent number of characters in each of our languages compared to 140 characters in English. These results are shown by the scale on the right y-axis for each set of box plots where the mean for English is set to 140 characters. Based on the UDHR outcomes, we find that 140-character worth of English content can be expressed in 35.53 Chinese characters or 55.71 Japanese characters. Similarly 140 characters worth of English content of a TED talk can be expressed in 43.61 Chinese characters or 56.70 Japanese characters. The differences between our two corpora are likely due to the nature of the content of each corpus, as the language used in the UDHR is much more formal and legalistic in comparison to the transcripts of TED talks.[3]

Beyond character measurements, we can also measure space more traditionally using the number of binary digits (or bits, eight of which form a byte) needed to store the same information content in different languages. The number of bits needed to store a character depends on the encoding scheme used. The most popular multilingual encoding scheme in use today is UTF-8, which uses a variable number of bytes to store a character. Almost all the characters used in English along with common punctuation marks require 8-bits (one byte) to store, while most characters used in Japanese and Chinese require 24-bits (three bytes) to store. Results from both our corpora (Figure 2) show that information in

[3]One limitation of using TED talks is that the vast majority of talks are given in English and other language transcripts are translations of this English source. There may be differences in how closely transcripts in each language follow the spoken dialogue, but we find that filler words (e.g., uh, er, and um) are rarely transcribed in any language, including English. Future work may consider developing a more balanced corpus by using movie subtitles with a variety of source languages.

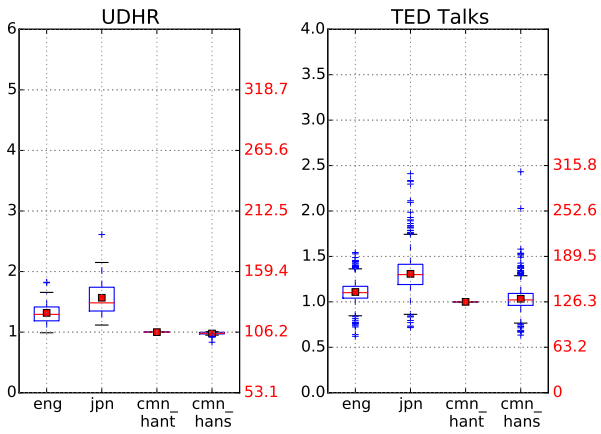

Figure 2: Relative ratio of bits (when the UTF-8 encoding scheme is used) required to express the same content when compared to Traditional Chinese (cmn_hant).

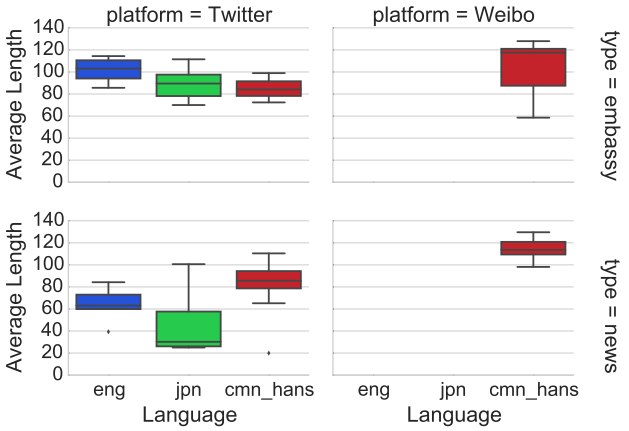

Figure 3: Length of microblog posts in characters (excluding URLs) in English (eng), Japanese (jpn) and Simplified Chinese (cmn_hans).

all our languages can be stored in similar amounts of space using UTF-8 encoding. While text in English requires more characters, the number of bits in UTF-8 required to store those characters is similar to the number of bits needed to store the same content in Chinese or Japanese.

We next turn to the questions of how much is typed and how much is said in microblog posts of different languages by different types of organizations. In answering these questions, we use the character calculations from the TED talk corpus. The number of characters used in a more universal measure that is directly apparent to the user with a larger effect upon the user experience.[4] Character limits are also the type of limits most users of microblogging platforms are familiar with since Twitter's limit is set this way. We use the measures calculated from the TED talk corpus rather than the UDHR corpus as a more conservative estimate of the differences between languages. Furthermore, the type of language used in microblog posts is likely closer to the language used in the TED talks than to the language used in the UDHR. Using Chinese with simplified characters (cmn_hans) as the baseline, we have the following ratios:

$$ratio(eng, cmn\_hans) = 3.21 \quad (4)$$

$$ratio(jpn, cmn\_hans) = 1.30 \quad (5)$$

3.2 How much is typed and posted: Organizations in action

We find that different organizations on Weibo and Twitter platforms post messages of slightly different character lengths in different languages.

[4]In contrast, the number of bits needed to store text depends on the character encoding used. Even within UTF-8, there are multiple ways to store accented characters and other characters such that two pieces of text that appear identical to the user could actually require a different number of bytes to be stored in UTF-8.

Figure 3 summarizes the overall outcomes of the microblog post length (in characters) with box plots. The left subplots show the results for Twitter, and the right subplots show the results for Weibo. The top subplots show the results for embassies, and the bottom subplots show the results for news organizations.

The average lengths of posts by English-language Twitter accounts (105 characters with URLs; 81 without URLs) are close to the "ideal length" of a tweet reported by marketing/engagement companies of 100 or 71–100 characters [17, 35]. For English, the embassy-type accounts on average have longer posts than the news-type accounts.

The average lengths of Japanese-language Twitter accounts show a large variance, particularly within embassy-type organizations. Although this observation should not be taken as conclusive because of the limited number of Japanese-language embassy accounts ($N = 18$), it is nonetheless indicative to see the relatively wide variation among the embassy-type accounts. In addition, within the news-type accounts, some Japanese-language tweets are shorter than both English and Chinese.

The average lengths of posts by Chinese-language Twitter accounts tend to be shorter than English-language posts for embassy-type organizations, whereas they tend to be longer than both English and Japanese posts for news-type organizations. The shortest average length of all the accounts in our dataset belongs to a Chinese-language Twitter account used by Falun Gong (screen name "djy_cn") to broadcast news in mainland China. Comparing Twitter and Weibo in Chinese, we find that the average lengths of posts by Chinese-language Weibo accounts tend to be longer than their Twitter counterparts.

3.3 How much is said: Comparing languages and organizations

Combining the two sets of findings above, this final subsection derives the amount of information content in posts in different languages by different types of organizations. Similar to Figure 3 in layout, Figure 4 shows the estimated information content using the Chinese language (with sim-

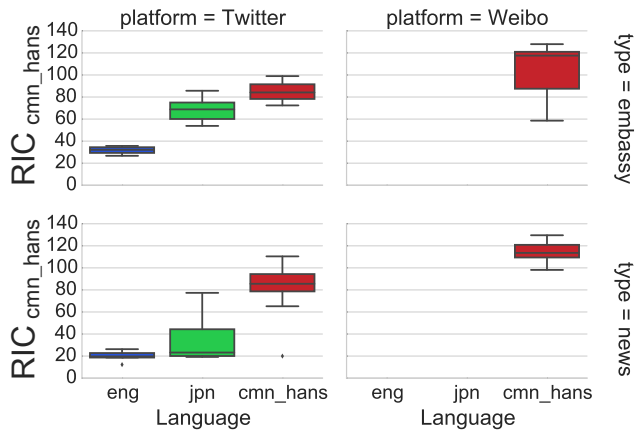

Figure 4: Relative information content (RIC) of microblog posts in English (eng), Japanese (jpn) and Simplified Chinese (cmn_hans). RIC is shown here as the equivalent number of Simplified Chinese characters (RIC_{cmn_hans}).

plified characters) as the baseline.

Effectively, what distinguishes Figure 4 from Figure 3 is that the number of characters in English and Japanese posts have been divided by the information content ratios found using the TED talk corpus (Equation 4 and Equation 5 respectively).

In contrast to the average length in characters, the comparisons of information content show a clear pattern. Most Chinese language microblog posts contain more information than either posts in Japanese or English. Furthermore, posts on Sina Weibo often contain even more content than posts on Twitter.

We see a larger variation among new organizations in comparison to embassies. Reflecting the variance in character lengths, the variance in information content per post among Japanese news organizations is quite large. By manually inspecting the data, we found that the shorter posts often included only a news headline while the longer posts often contained a short summary of the news story.

Across the three languages, the findings suggest a general pattern where Chinese-language posts contain more information per post than either Japanese-language or English-language posts. This difference is strongest for embassies (the top two subplots in Figure 4). For the news organizations, a higher degree of overlap exists. For example, there are Chinese-language Twitter accounts such as Falun Gong (shown as an outlier dot in the figure) posting short messages. The variation in Japanese-language accounts means that some Japanese-language accounts post messages with more information content than some Chinese-language accounts, but other Japanese-language accounts post messages with an amount of information content that is similar to many English-language accounts.

Chinese-language posts are generally longer and contain more information on Weibo than Chinese-language posts on Twitter. Confusingly, we found that some Weibo posts contained more than 140 characters whereas no Twitter posts did. As explained previously, based on further manual examination it appears that Sina Weibo enforces a byte-length limit using a variable-length character encoding such as GBK. We found that on Sina Weibo that we were able to post messages using up to 140 Chinese characters or 280 English characters. This difference in characters available to users is an important distinction and one that our data show many Weibo users make use of as the inclusion of URLs and other English/Latin characters leaves more space for more text (of any language) on Weibo in comparison to Twitter.

4. DISCUSSIONS AND CONCLUSIONS

Based on a corpus linguistic approach with open, crowdsourced translation data, we have updated the cross-lingual ratios of information content between English, Chinese, and Japanese. Then we have applied the results to measure empirically how much can be said and how much is actually said in microblog posts of different languages on Twitter and Sina Weibo.

The construction of relative information content measures allows researchers to move beyond length-based comparisons for microblogging to consider other content. Wikipedia, for instance, maintains a list of the one thousand most important articles and compares their lengths across language editions using "language weights" (relative to English) [22]. Our findings analyzing both the UDHR and TED talks show the impact of corpus selection for determining such weights, and calculating accurate weights will require a parallel corpus with a high degree of similarity to the type of text commonly found in Wikipedia.

We find differences in microblogging activity most strongly by platform and language, but also by organization type. In general, English-language posts use more characters than either Japanese- or Chinese-language posts on Twitter. However, once the information content per character of each language is taken into account, the relative information content per post shows that English-language posts actually contain less information per post than either Japanese- or Chinese-language posts. As a consequence, the very definition of what constitutes "micro" on each platform differs by language, and this suggests that the user experience of microblogging platforms may differ greatly between languages. We further find that information content differs between Twitter and Weibo, with posts on Weibo generally containing more information than posts on Twitter. This platform difference is likely a consequence of how the platforms enforce their length limits. Twitter enforces a limit of 140 characters without regard to the storage requirements of the characters, while Weibo enforces a byte-limit.

Our work thus adds to the existing scholarship [e.g., 15, 24, 27] showing how the user experience of the same platform differs for users writing content in different languages. While character or byte limits impose a superficially similar measure of length across languages, we find considerable difference in the actual amount of information content available and commonly included within microblog posts written in different languages. Hence such a superficial measure of length actually results in wide user-experience differences. These user-experience differences impact both content consumption and content creation as demonstrated by the literature on the impact of length restrictions on SMS messages [e.g., 10, 21]. Our findings have important ramifications for efforts to translate content across languages: content that fits within one Japanese- or Chinese-language post may not

fit within one English-language post when character limits are imposed. Our findings also suggest that if the "ideal" length to maximize engagement sought after by marketing companies does exist, it is almost certainly language dependent.

The reliance on certain language-dependent properties of information content is bound to raise a fundamental question about how such design parameters can be uncritically applied to other languages and platforms. It also challenges the perceived wisdom such as the "ideal length" of tweets on the grounds that these previous findings cannot be generalized straightforwardly to other languages and/or platforms. Our findings show clear differences between the information content commonly contained in Japanese-, Chinese-, and English-language posts.

From the perspective of multilingual Internet or Internet linguistics, digital support for the main East Asian languages is an important milestone for the internationalization of the Internet. As put by Nakayama Shigeru, a major East Asian Science Technology and Society scholar [31]:

> East Asians are accustomed to dealing with a multi-byte system, in contrast to Western mono-byte reductionist culture. It may be that in the future our multi-byte culture will prove advantageous for dealing with complex systems (p. 12).

Although one does not have to agree with Shigeru's criticism of "Western mono-byte reductionist culture," the findings here do suggest that platform designers and researchers need to carefully analyze what settings and assumptions may be language-specific.

By using open and freely available user-generated translation data as parallel corpora and by collecting both Twitter and Weibo posts in three languages, this research has investigated the language-specific effects of character limits on microblogging. It is expected that more systematic measurement and more linguistically diverse data sets will help both researchers and designers reexamine some of the designs and practices that are in reality language-dependent and/or language-biased and thereby find ways to account for them and develop better designs and research that are language-aware and/or language-neutral. Further research is necessary in this area to expand our knowledge of internationalization in Web Science and Internet research as well as cross-cultural Human-Computer Interaction (HCI).

Our work has been the first to investigate character lengths, byte lengths, and information content across two major microblogging platforms. Future work will build upon the work presented here to increase sample sizes, language coverage, and the types of users included. The use of less formal, more natural parallel corpora (i.e., TED talks) rather than formal legalistic prose for understanding informal conversation on Twitter and Weibo is also an important contribution. As the efforts to build parallel corpora from Twitter and other user-generated content platforms [e.g., 4, 25, 32] and from the Web more generally [e.g., 18, 28] improve and organizations such as Meedan and Global Voices continue facilitating the human translation of user-generated content, we will have additional tools and corpora through which to examine the impact of length and other constraints on Internet-mediated communication.

Acknowledgments

This research was supported in part by the University of Oxford's ESRC Impact Acceleration Account and Higher Education Innovation Fund (HEIF) allocation. The authors would like to thank our anonymous reviewers for their helpful comments.

5. REFERENCES

[1] 3rd Generation Partnership Project (3GPP). Digital cellular telecommunications system (phase 2+); universal mobile telecommunications system (umts); technical realization of the short message service (sms) (3gpp ts 23.040 version 12.2.0 release 12). Technical report, 2014. `http://www.etsi.org/deliver/etsi_ts/123000_123099/123040/12.02.00_60/ts_123040v120200p.pdf`.

[2] Y. Chan. Microblogs reshape news in China. China Media Project, Oct. 2010. `http://cmp.hku.hk/2010/10/12/8021/`.

[3] J. Constine. Facebook ups character limit to 60,000, Google+'s is still bigger. 2011. `http://techcrunch.com/2011/11/30/status-update-character-limit/`.

[4] J. Dalton. Useful data from "pointless babble": Discovering sentence-aligned parallel texts from Twitter streams. Master's thesis, Department of Computer Science, University of Oxford, 2012.

[5] L. Dugan. 140 Characters On Chinese Twitter Is More Like 500 Characters On Twitter.com. AllTwitter, July 2011. `http://www.mediabistro.com/alltwitter/140-characters-on-chinese-twitter-is-more-like-500-characters-on-twitter-com_b11951`.

[6] M. Ebner and M. Schiefner. Microblogging - more than fun. In *Proceedings of the IADIS International Conference on Mobile Learning*, pages 155–159. IADIS, 2008.

[7] K.-w. Fu, C.-h. Chan, and M. Chau. Assessing censorship on microblogs in China: Discriminatory keyword analysis and impact evaluation of the "Real Name Registration" Policy. *IEEE Internet Computing*, 17(3):42–50, 2013.

[8] K.-w. Fu and M. Chau. Reality check for the Chinese microblog space: A random sampling approach. *PLOS ONE*, 8(3):e58356, 2013.

[9] C. Gayomali. The text message turns 20: A brief history of SMS, Dec. 2012. `http://theweek.com/articles/469869/text-message-turns-20-brief-history-sms`.

[10] R. E. Grinter and M. A. Eldridge. y do tngrs luv 2 txt msg? In *Proceedings of the Seventh European Conference on Computer-Supported Cooperative Work*, ECSCW 2001, pages 219–238. Springer Netherlands, 2001.

[11] S. A. Hale. Impact of platform design on cross-language information exchange. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems Extended Abstracts*, CHI EA '12, pages 1363–1368, New York, NY, USA, 2012. ACM.

[12] S. A. Hale. Global connectivity and multilinguals in the Twitter network. In *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems, CHI '14, pages 833–842, New York, NY, USA, 2014. ACM.

[13] D. Hewitt. Has Weibo really changed China? BBC, July 2012. http://www.bbc.co.uk/news/magazine-18887804.

[14] F. Hillebrand. Short message and data services, 2002.

[15] L. Hong, G. Convertino, and E. Chi. Language matters in Twitter: A large scale study. In *International AAAI Conference on Weblogs and Social Media*, pages 518–521, 2011.

[16] A. Java, X. Song, T. Finin, and B. Tseng. Why we Twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[17] K. Lee. The proven ideal length of every tweet, Facebook post, and headline online. Fast Company, Apr. 2014. http://www.fastcompany.com/3028656/work-smart/the-proven-ideal-length-of-every-tweet-facebook-post-and-headline-online.

[18] B. Li and J. Liu. Mining Chinese-English parallel corpora from the Web. In *In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*, pages 847–852, Hyderabad, India, 2008. Citeseer.

[19] H.-T. Liao. What do Chinese-language microblog users do with Baidu Baike and Chinese Wikipedia? a case study of information engagement. In *Proceedings of The International Symposium on Open Collaboration*, OpenSym '14, pages 23:1–23:10, New York, NY, USA, 2014. ACM.

[20] M. Liberman. Language Log: One world, how many bytes? Language Log, Aug. 2005. http://itre.cis.upenn.edu/~myl/languagelog/archives/002379.html.

[21] R. Ling and N. S. Baron. Text messaging and IM linguistic comparison of American college data. *Journal of Language and Social Psychology*, 26(3):291–298, Sept. 2007.

[22] Meta-Wiki. List of Wikipedias by sample of articles, 2015. http://meta.wikimedia.org/wiki/List_of_Wikipedias_by_sample_of_articles.

[23] M. Milian. Why text messages are limited to 160 characters. LA Times Blogs, May 2009. http://latimesblogs.latimes.com/technology/2009/05/invented-text-messaging.html.

[24] A. X. Mina. Batman, pandaman and the blind man: A case study in social change memes and Internet censorship in China. *Journal of Visual Culture*, 13(3):359–375, 2014.

[25] M. Mohammadi and N. GhasemAghaee. Building bilingual parallel corpora based on Wikipedia. In *Computer Engineering and Applications (ICCEA), 2010 Second International Conference on*, volume 2, pages 264–268, March 2010.

[26] G. Neubig and K. Duh. How much is said in a tweet? A multilingual, information-theoretic perspective. In *AAAI Spring Symposium on Analyzing Microtext*, 2013.

[27] U. Pfeil, P. Zaphiris, and C. S. Ang. Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113, 2006.

[28] P. Resnik and N. A. Smith. The Web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380, Sept. 2003.

[29] R. J. Rosen. How much can you say in 140 characters? A lot, if you speak Japanese. The Atlantic, Sept. 2011. http://www.theatlantic.com/technology/archive/2011/09/how-much-can-you-say-in-140-characters-a-lot-if-you-speak-japanese/245199/.

[30] B. Ruby. Twitter versus Weibo: What You Need To Know. The Fearless Group, Nov. 2012. http://thefearlessgroup.com/twitter-versus-weibo-what-you-need-to-know/.

[31] N. C. Shigeru. The digital revolution and East Asian science. In A. K. L. Chan, G. K. Clancey, and H.-C. Loy, editors, *Historical Perspectives on East Asian Science, Technology, and Medicine*. World Scientific, Singapore, 2001.

[32] J. R. Smith, C. Quirk, and K. Toutanova. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 403–411, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[33] B. Summers. What's the equivalent of Twitter's 140 character limit for non-Latin character sets?, Feb. 2010. http://bens.me.uk/2010/twitter-charset-experiment.

[34] H. Sun. *Cross-Cultural Technology Design: Creating Culture-Sensitive Technology for Local Users*. Oxford University Press, Oxford, Mar. 2012.

[35] Track Social. Track social blog: Optimizing Twitter engagement – part 3: Tweet length, Oct. 2012. http://tracksocial.com/blog/2012/10/optimizing-twitter-engagement-part-3-tweet-length/.

[36] Twitter. Best practices: Be the best at what, when and how you tweet. https://web.archive.org/web/20131020033342/https://business.twitter.com/best-practices.

[37] Twitter. Character counting. https://dev.twitter.com/overview/api/counting-characters.

[38] United Nations. The World's Most Translated Document. Human Rights Day, Dec. 2007. http://www.un.org/en/events/humanrightsday/2007/worldtransdoc.shtml.