Readability of Twitter Tweets for Second Language Learners

Patrick JACOB and Alexandra L. UITDENBOGERD
RMIT University - School of Science
124 La Trobe St
Melbourne VIC 3000
patrick.jacob@rocketmail.com,sandra.uitdenbogerd@rmit.edu.au

Abstract

Optimal language acquisition via reading requires the learners to read slightly above their current language skill level. Identifying material at the right level is the essential role of automatic readability measurement. Short message platforms such as Twitter offer the opportunity for language practice while reading about current topics and engaging in conversation in small doses, and can be filtered according to linguistic criteria to suit the learner. In this research, we explore how readable tweets are for English language learners and which factors contribute to their readability. With participants from six language groups, we collected 14,659 data points, each representing a tweet from a pool of 4100 tweets, and a judgement of perceived readability. Traditional readability measures and features failed on the data-set, but demographic data showed that judgements were largely genuine and reflected reported language skill, which is consistent with other recent studies. We report on the properties of the data set and implications for future research.

1 Credits

We thank Klerke et al. (2016) for the Twitter corpus from their unpublished research.

2 Introduction

Since the first half of the twentieth century researchers have analysed texts to determine their *readability*, that is, how easy the text is to read and comprehend, often expressed as levels of linguistic education, knowledge, age or experience. The findings around readability have been applied to education for selecting appropriate reading material for students, in communication with governmental bodies to reach a higher number of citizens (Temnikova et al., 2015) and in marketing/public relations of companies to increase the reach of their materials (Risius and Pape, 2016).

While there have been many studies on readability of regular text for students and foreign language learners, the same is not true for the microblog text genre. Twitter is a popular platform for reading current topics and engaging in social interaction, providing a cross-cultural, cross-interest, and cross-language platform for reading social media posts of up to 280 characters in length, and a filtered feed has potential as a source of regular reading material for learners. While the Flesch readability of English language tweets has been analysed to discover demographic trends and to compare them to other modern text genres (Davenport and DeLine, 2014), and judgements of tweet clarity for emergency communication has been researched (Temnikova et al., 2015), to our knowledge tweets have not been studied in relation to English as an Additional Language (EAL, a term that recognises that it may be a third language, for example).

Our aim was to extend readability research to tweets for EAL learners. Tweets are different from ordinary text due to their short length, hashtags, *mentions* (user identifiers preceded by an @ symbol), links, and other non-standard text tokens that they contain. This poses challenges for traditional readability formulae, which assume regular text, as found in books and periodicals. This study evaluates the applicability of readability formulae and the influence of the unique expressions used in tweets such as hashtags, mentions and links. The goal was to find predictors that increase accuracy in classifying Twitter tweets to language reading levels, which will assist users to find more appropriate material for their reading abilities and aid institutions to adjust their published tweets for foreign language reader target audiences.

3 Related Literature

3.1 Classic Studies

Text readability has been researched since early last century, and produced the widely used Flesch Reading Ease formula (Flesch, 1948).

$$206.835 - 1.015\left(\frac{words}{sentences}\right) - 84.6\left(\frac{syllables}{words}\right)$$

The Flesch formula shows an inverse relationship between readability and the number of syllables per word (lexical complexity) and the number of words per sentence (grammatical complexity). This simple measure has become a standard for text analysis in other fields of research, and is often used as a baseline for readability research, hence we include it in our study. Dale-Chall (1948) is another user-derived readability measure, based on children with English as a first language:

$$0.1579\left(\frac{difficult\ words}{words} * 100\right) + 0.0496\left(\frac{words}{sentences}\right)$$

Dale and Chall's research determined that the percentage of difficult words in a given text and the number of words per sentence influence readability. This formula assumes every word not on a list of 3000 words a fourth-grade American student should be familiar with is difficult. It would be interesting to see the interaction between the Dale-Chall formula and research based on the findings of Uitdenbogerd (2005), which show that cognates (words that are same or very similar between the native language and the foreign language of study) influence the understanding of sentences for students of foreign languages.

Most readability measures and indexes are only considered valid for text samples with a minimum number of words or sentences (Collins-Thompson and Callan, 2004; Homan et al., 1994), and therefore not intended for typical tweet text. However, we include the above formulae and related classic readability features in this initial study.

3.2 Twitter-related Research

Davenport and DeLine (2014) studied the readability of a corpus of 17.4 Million tweets. They modified the Flesch formula by treating each tweet as a single sentence, due to their brevity and unconventional punctuation. This approach may no longer be adequate for a Twitter corpus, given the new character limit of 280 characters for tweets.

Temnikova et al. (2015) analysed the text difficulty of emergency messages on social media including Twitter. They used crowd-sourcing (CrowdFlower) to present a questionnaire of 500 tweets to participants, who rated them as one of *very clear*, *needs improvement*, or *very unclear*. Additionally, participants could suggest how to write a more understandable version of the tweet. Amongst the resulting recommendations are to use easy vocabulary and short complete sentences, exclude mentions, and minimise hashtag use. Even though the resulting recommendations appear to be valid, it is unclear what the background of the participants was, which can impact how text is perceived. In contrast, for our study, we selected and recorded the background of participants from specific populations.

4 Experiment Design

There are generally two types of research design for predicting readability. The first models reading difficulty using data collected from human participants. The readability measure by Kincaid et al. (1975) is one example of this. They invited 531 participants from two navy bases in the US to read from a set of eighteen passages of training manuals. The task was to answer questions about the manuals by filling in missing words (Cloze test). From the results, Kincaid et al. deduced the formula to predict the reading grade level for navy personnel. The advantage of this approach is that the collected data and resulting model represents the genuine user experience of text difficulty. The main challenge is obtaining sufficient data from the target user population for analysis.

The second research method, which has become prominent in NLP communities, uses large corpora of text samples that have been labelled by experts or publishers, to train machine learning models. One example is the research of François and Fairon (2012), who trained a machine learning algorithm with a text corpus labelled according to the levels of the Common European Framework of Reference for Languages (CEFR), to model the readability of French text for second language learners (François and Fairon, 2012). This approach allows modern classifiers to be trained on large data-sets of features. However, as has been confirmed by Vajjala and Lucic (2019), expert or

publisher labels of text are a poor substitute for genuine user experience, and even the choice of method of measuring the reading experience can lead to large differences in results. This echoes the results found elsewhere in usability research (Jeffries and Desurvire, 1992).

There was no Twitter corpus annotated with difficulty levels available, hence our research design consisted of a user study of tweet readability, specifically for people with English as an additional language. Our approach has the added advantage of reflecting the user experience of language learners matching the demographics of the participants. Participants completed a questionnaire that collected demographic data and reading difficulty judgements of a set of tweets.

4.1 Participant Recruitment

Wilson VanVoorhis and Morgan (2007) recommends that with more than six predictors, to have at least ten participants per predictor. With 10-15 predictors from the survey (such as age groups, Twitter affinity, education levels) and text features from the tweets (such as the number of syllables, characters or Hashtags), we needed at least 150 participants per language. To account for contradicting, invalid or otherwise wrong responses that would need to be discarded from the corpus, we increased the target number of recruits to 200.

We tried to recruit 200 native speakers from each of the six target languages of our study (Spanish, Portuguese, German, Dutch, Cantonese, and Mandarin) via the crowd-sourcing platform Figure Eight[1]. The actual questionnaire was hosted on Qualtrics[2], a specialised website for conducting questionnaires. A participant would be forwarded to the Qualtrics questionnaire via a link once they accepted the survey questionnaire.

4.2 Twitter Corpus Collection

The Twitter corpus we used was merged from two Twitter corpora: one corpus from unpublished research by Klerke et al. (2016); and a larger corpus initially containing 6,000 randomly captured tweets using the Twitter Stream Application Developer Interface.

The second corpus was captured in August 2018 directly from the Twitter stream, utilising the `tweepy` python library[3], which allows searching for a specified number of tweets that contain defined keywords. We used both functions to search for about 400 English language tweets for each first language, containing at least one word from a list of cognates of that language. This ensured that the tweet corpus contained a minimum number of cognates from each language. Due to specific post-collection steps that lowered the final corpus of tweets, more tweets were collected than needed.

Original: *People Swea They KNOW E V E R Y T H I N G Bhou Me Bhuh They Dont Know NOTHING Bhuu my Name*
Corrected: *People swear they know everything about me, but they don't know nothing but my name*

Figure 1: An example of tweet simplification

Tweets were filtered for offensive content, using an automatic profanity check, followed by a manual process by the researchers to filter any remaining offensive tweets. Lastly, we filtered and deleted duplicates (such as retweets) leaving the entire corpus at 4700 tweets, commencing with 873 from the Klerke corpus. The first 4000 were used for the survey.

Due to platform limitations we broke the survey up into five surveys for each language: four of 1000 tweets and one of 100 tweets used for further validity checking and analysis. The 100-tweet survey consisted of tweets originally containing colloquialisms and/or social media features, such as emojis, which were manually selected from the pool of 4000. The tweets were stripped of emojis, hashtags, mentions, and repetitious content; spelling corrected, and the text adjusted in other ways to standardise it (for example, see Figure 1). It was used to test the questionnaire setup prior to releasing the main surveys.

4.3 Questionnaire

To avoid reading fatigue and to stay within the budget, each person made 20 judgements. This approach should have resulted in six judgements per question for 4000 tweets. That is, for each tweet, we would have at least two human judgements from each language family group. Using the Qualtrics randomisation function, the tweet questions were selected randomly from the pool of 4100 tweets to minimise ordering effects. To ensure an even distribution of judgements, each tweet was presented to at least one participant before any were shown a second time.

Participants were asked for their age, gender, country, education and foreign language knowl-

[1] https://www.figure-eight.com
[2] https://www.qualtrics.com/au
[3] https://www.tweepy.org

edge, to assist in providing context for the ground truth collected, as well as to capture potential confounding variables known to influence vocabulary knowledge. We then presented the participants with 20 tweets for them to judge according to reading difficulty. Participants were to position a slider on a scale from 1 (very difficult) to 10 (very easy) representing their perception of the tweet's readability, as shown in Figure 2.

The last task for the participant was to answer a short translation task to confirm the participant does indeed speak their stated first language. The translation question was based on common proverbs in the participant's native language, which they needed to translate from English to their native language. This had the advantage that it was a relatively easy task, since proverbs are usually widely known, but allowed us to evaluate if the participant speaks the claimed language.

Using the IP range of specific countries, we restricted the survey job to specific language speakers in countries where they predominately or officially spoke that language. This way we had another layer to ensure we would only recruit the right target participants.

4.4 Survey Execution and Outcome

For the 100-tweet test surveys we lowered the number of tweets per job from twenty to ten. After seeing that target participation was reached for three languages (Spanish, Portuguese and German) we released all other jobs, which were kept open for about a month. Table 1 shows that Spanish, Portuguese and German participants were most active, while Chinese and Dutch-speaking countries had much lower participation. In the case of Dutch-targeted jobs, someone hacked the survey and exhausted the available budget, leaving us with few judgements for Dutch speakers.

5 Data Restructuring and Cleansing

Data cleansing prior to analysis consisted of the following steps:

- Transposing the data columns into a format suitable for analysis
- Harmonising the contents of several columns such as country of origin or languages.
- Matching and unifying the columns about educations levels.
- Deleting rows with failed validation questions.

Table 1 shows the final data set size for each language after the data cleansing steps were finished, .

Survey	Number of participants	Number of data points
Spanish	258	4188
Portuguese	233	4187
German	240	4179
Dutch	55	928
Mandarin	35	547
Cantonese	44	630
Total	865	14659

Table 1: Number of data points after cleaning

6 Descriptive Statistics

When visualising the judgement data as a histogram (see Figure 3) it shows an exponential distribution from very difficult to very easy perceived tweets. Fitting a line to the log of the number of judgements at each rating level has an R^2 of 0.97. Thus most tweets were evaluated as 10 (very easy to read and understand) by participants.

6.1 Twitter Use

When looking at the average ratings shown in Table 2, it can be seen that the more time someone spends with Twitter, the easier it is for participants to read tweets. Participants who used Twitter daily or weekly rated the tweets at 8.39 on average, while participants that never used Twitter averaged 7.99. Presumably frequent Twitter users are more accustomed to the linguistic conventions of Twitter and find it easier to understand tweets. This would partially explain why the majority of tweets are rated 10, as the majority of participants were heavy Twitter users.

Twitter usage	M	SD	Sample Size
Daily	8.39	1.85	6988
Weekly	8.39	1.96	3409
Occasionally	8.14	2.02	3156
Never	7.99	2.07	1109

Table 2: Mean and standard deviation of tweet readability judgements across Twitter usage groups

6.2 Education

Formal school and language education had a strong influence on the judgements in the data

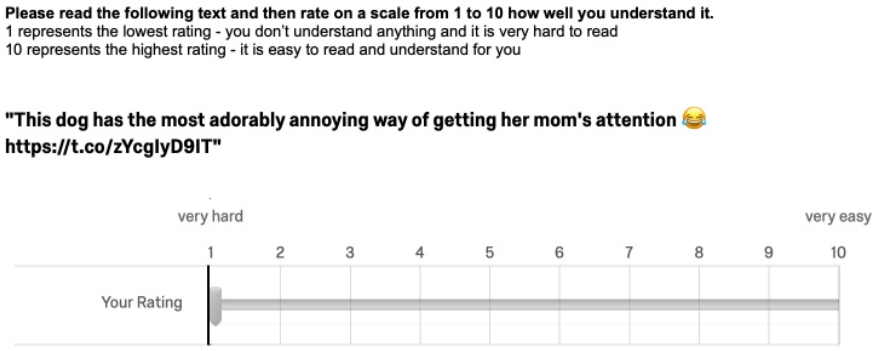

Figure 2: Example tweet question including slider.

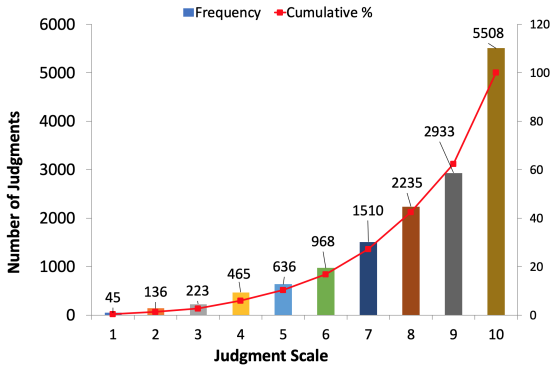

Figure 3: Histogram of judgements

(see Figure 4). Participants without any formal education didn't rate any tweets as 10, while the group of PhD graduates have the highest fraction of tweets rated 10. PhD graduates judged tweets as 9.07 on average, whereas participants without formal education rated their tweets on average at 7.25.

CEFR Level	M	SD	Sample Size
A1	9.35	1.0	160
A2	7.8	2.08	720
B1	8.46	1.77	1311
B2	8.52	1.66	1090
C1	8.64	1.62	547
C2	8.74	1.98	798

Table 3: Average judgement by CEFR Level

During data cleansing, we mapped all reported English education levels to the CEFR standard, which has levels in increasing order of skill, A1, A2, B1, B2, C1 and C2 respectively. This mapping was possible for 4523 data points, which represents 30% of all judgements. Our data shows that the higher the English education, the more likely the tweets are judged higher. The average of A2 participants is 7.8 (30% of tweets given a 10), while the average of the C2 group is 8.74 (65% of judgements being 10) and average ratings increase monotonically between those two levels. The exception is A1, which had an average of 9.35. This could be due to a Dunning-Kruger effect, in which those with minimal knowledge of a subject have a disproportionately high opinion of their knowledge, a problem with the CEFR mapping at the A1 end, or randomly assigned tweets coincidentally being easier to read. It should also be noted that there were only 160 A1-based judgements, whereas all other language groups had at least 547. Those with A1 level English or less are likely to have found the user interface itself challenging, let alone the tweets they were allocated, which may have impacted their participation, resulting in a high proportion of "false beginners" in the cohort.

No. of add. Lang.	M	SD	Sample Size
0	8.16	2.03	8105
1	8.43	1.84	5330
2	8.69	1.55	848
3	8.90	1.63	297
6	9.08	1.01	79

Table 4: Average judgement by additional languages spoken

We also captured any additional languages participants spoke besides their native language and English. Table 4 shows that the more languages a person spoke, the higher the average rating per tweet. The population of people speaking more than one additional language is relatively small, but so is the standard deviation. It is likely that

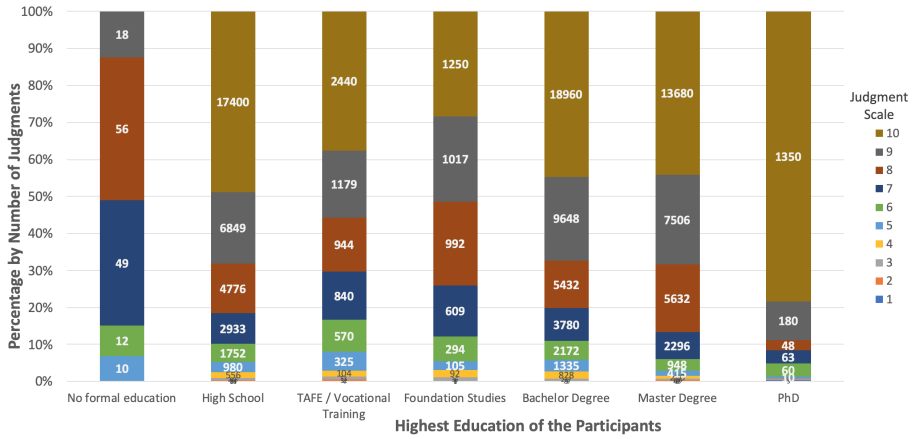

Figure 4: Chart of judgements by education level

broader language knowledge improves the reading capabilities of unusual text such as tweets.

6.3 Twitter-specific Text Features

Twitter is a social media platform, where additional features are used to graphically express emotions and other items (*emojis*); or connect with other users (*mentions*), tweets (*hashtags*) or websites in and outside of Twitter (*links*). We look at each of these features below.

Emojis Emojis are ideograms used in messaging, including stylised facial expressions for displaying emotions, places, animals, food, and flags, among other objects. For the large data set, tweets with 0, 1, 2, 3 and >3 emojis respectively all had ratings between 8.17 and 8.51 with no obvious trend, and standard deviations from 1.84 to 1.98. The emojis did not seem to influence the judgements.

We also analysed a subset (40) of the modified tweets from the small test set from which emojis had been stripped. The average judgements of tweets with emojis removed (8.18, $n = 246$) was lower than that of the original tweets ($M = 8.33$, $n = 93$) . Due to the universal understanding of emojis across languages, they *might* increase readability, or their removal from tweets may take essential semantic content away. However, the difference in means is small, the variability high, and the tweets themselves were not randomly selected, so strong conclusions cannot be drawn at this stage.

Hashtags Hashtags are used as metadata tags to reference themes or content and make them easily findable within and across social media platforms.

Hashtags per Tweet	Count	M	SD
>1	1215	8.14	2.01
1	1812	8.24	1.99
0	11632	8.33	1.92

Table 5: Mean and standard deviation of judgements according to number of hashtags

The question is if they influence the readability of tweets, since they are often composed of joined and abbreviated words, for example, *#muppetgovernment* or *#ImACeleb*. In our corpus the number of hashtags present ranged from zero to twenty, but with very few containing more than 3 hashtags, and no obvious trend was observed as hashtags increased without binning. As with emojis, the subset of 29 modified tweets stripped of hashtags was judged less readable on average (8.07, $n = 195$) than the original ones (8.43, $n = 61$). A reverse trend was found in the larger data-set (see Table 5), with minimal overlap of confidence intervals, indicating confidence in the estimate of the population mean. However, differences in the mean are much smaller than those of the standard deviation, so hashtags are not strong predictors of readability.

Mentions Mentions use the @-sign to refer to other users on Twitter, and like hashtags, are often used on social media, typically either at the beginning or end of tweets. Twitter does not count mentions in the character limit but only allows up to 50 mentions per tweet.

We used the test subset (22) to compare tweets that are stripped of mentions against those with mentions. On average, the judgement with modified tweets is 8.19 ($N = 156$), while the ones with

mentions lie at 7.89 ($N = 72$). These numbers indicate that mentions decrease readability.

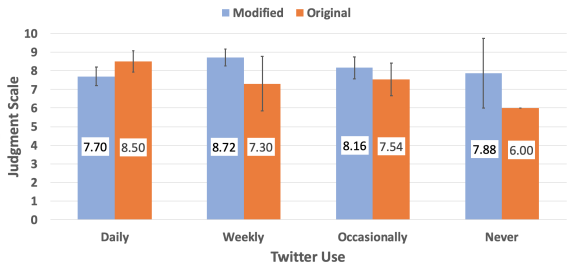

Figure 5: Chart of average judgements by mention broken down by Twitter use. Error bars are 95% confidence intervals.

Interestingly, when broken down by Twitter use, daily Twitter use led to tweets with mentions being rated higher than those without. The opposite was true for those who used Twitter weekly or less frequently. This behaviour could mean that frequent users are better able to filter or appropriately process mentions when reading.

Links Links are often used on Twitter to refer to other resources on the internet, such as news articles or videos. The links are often abbreviated to save space (for example, *https://t.co/hle8l0AO1i*). We found no evidence that links influence judgements of tweets, whether we used the number of links or their length.

6.4 Readability Measures

We analysed the relationship between different text features and judgements, such as the number of characters per word, number of syllables per word, and sentence length. Most of them had negligible impact, except the total number of characters or words seems to show a trend on average that the more words, the lower the rating, but since they feed into readability measures, we would like to point out a few findings with traditional readability formulae.

Flesch Reading Ease We compared judgements with Flesch scores, grouping scores into bins, which showed a peak at RE $\in$ [77,81], indicating most tweets were in the fairly easy to easy range. A slight upward trend was observed, indicating weak agreement between RE and judgements.

Some Flesch scores were extremely negative, due to the sentence length and frequent use of words with high syllable counts. For example the following tweet is one sentence long, with 30 Syllables and eight words with a Flesch Score of -118.53.

#FollowMikeaveli #FollowMikeaveli #FollowMikeaveli #FollowMikeaveli NO QUESTIONS JUST FOLLOW.

While this tweet had a very negative score, meaning very difficult, its three ratings were 10, 10 and 7. We also tried the Flesch-Kincaid formula, which had similar trends.

Dale-Chall The feature that is unique to Dale Chall's formula is the number of difficult words, being all words not in a list of 3000 easy words. Our data shows that on average, tweets with a higher number of difficult words are judged more difficult. The Dale-Chall formula however, shows the opposite trend. That is, the harder the tweet according to the Dale-Chall score, the easier it was judged by participants. Additionally, we did an analysis and exchanged cognates for "easy words" in the formula to see if this would have any effect. The trend reversed to the expected direction, but was again weak. A more nuanced approach is probably needed, with high frequency words from the list retained and combined with cognates. This will be explored in future work.

6.5 Correlation

Using both Pearson and Spearman correlations, we calculated correlation matrices between all columns and features. We used both formulations, as Pearson calculates the linear relationship between two variables, while Spearman evaluates the monotonic relationship, which is more appropriate for ordinal data or not entirely linear data. (See Table 6.)

No single feature has a strong relationship to the judgements. The range is between negative and positive 9.6%, which is quite low. Correlation between native languages was also low, regardless of language similarity. This could mean that readability is different for each language. The highest positive Pearson correlation, and second highest Spearman, is the number of additional languages a participant speaks. Education and English level are also highly placed, confirming the previous finding that education or language skills have a stronger relationship with readability than the content itself. Twitter-specific features like the number of emojis and hashtags have little relationship with the judgements, the strongest being for mentions (Spearman -5.1%). In general, we find

Features	Pearson	Spearman
number_of_further_languages	9.6%	8.3%
english_level	7.2%	6.2%
twitter_usage	6.4%	4.8%
education	4.8%	4.4%
flesch_kincaid_twitter_adjusted	4.2%	5.3%
flesch_1948	3.2%	3.3%
percentage_cognates_per_Tweet	1.8%	0.2%
dale_chall	1.2%	-0.2%
number_of_Emojis	0.4%	0.9%
average_length_links	-0.5%	-1.3%
number_of_links	-0.6%	-1.4%
number_of_sentence	-1.6%	-3.3%
number_of_hashtags	-2.3%	-2.9%
number_of_cognates	-2.5%	-3.2%
number_of_mentions	-2.7%	-5.1%
cognates_dale_chall	-3.1%	-2.1%
flesch_kincaid	-4.5%	-5.0%
number_difficult_words	-5.3%	-8.2%
number_of_words	-5.5%	-8.5%
number_of_syllables	-5.8%	-8.6%
number_of_characters	-6.0%	-9.2%

Table 6: Pearson and Spearman rank correlation between judgements and features.

that demographic data are stronger predictors than text features.

6.6 Confidence in Results

Our correlation matrix showed that no feature has a strong correlation to the judgements. It made us question whether the results were trustworthy or whether the participants put in any effort. While we had a validation question for each participant to check if they spoke the native language they claim, we did not implement a similar question to measure sincerity in answering. However, we have some indication that participants answered thoughtfully. First, the slider for tweets was initially set to 1, representing very hard to read and understand, but most of the tweets were rated 10. It means the participants moved the bar to provide their response. Second, the test subset (17) we manipulated to more straightforward language (see, for example, Figure 1), had an average judgement of 8.35 ($n = 99$) compared to the original average of 7.55 ($n = 56$), meaning simplified ones were judged as easier to read. These results lower our doubts about the sincerity of the answers by the participants.

7 Future Work

The features we extracted have a low correlation to the judgements. However, these are not the only features that can be extracted. We saw that uncommon or incorrect words have an effect on readability, therefore, constructing a measure of the severity of incorrectness might show stronger correlation than we currently have. Other possible features could be a percentile of non-lexical words, presence of particular grammatical terms or frequency of named entities to name a few. From the extracted features, emojis and mentions, while not showing high correlation themselves, may influence judgements when isolated and compared to tweets stripped of them. We are also yet to explore the use of features such as perplexity. Our machine learning results using further features will be reported elsewhere.

A difficulty with the current data set is that the majority of judgements are at the maximum of the scale, indicating a mismatch between participants and text. A new experiment that selects more homogeneous participant groups based on confounding variables such as age, Twitter usage, English levels and education may be more successful. Obtaining more judgements per tweet would allow more conclusions about user perceptions.

8 Conclusion

We started this research by asking what influences the readability of English tweets for foreign language speakers?

We designed and executed a survey on a crowdsourcing platform where 865 participants made 10–20 readability judgements from a pool of 4100 tweets. It did not produce the results we expected, as all features showed a low correlation ($\leq 9.6\%$) to the judgements. These features included traditional readability formulae and their components, which in other studies correlate well with user judgements (for example, Uitdenbogerd (2005) achieved 9-85% correlation for traditional readability features and formulae). This study revealed that traditional readability formulae do not work well on tweets. Another observation we made is that some demographic data had stronger predictive power than the text features themselves. For example, English skill level, number of languages known besides English, and the native language showed the highest correlation out of the available features.

As for what makes it hard or easy to read tweets, we do not have a definitive answer, but our research points in the following directions. Slang, wrongly written and uncommon words seem to lower the readability. The number of words

or characters and readability formulae have limited predictive value on the readability. From the Twitter-related features, emojis may improve readability, while using mentions and hashtags diminish it for those less familiar with tweets.

All these insights leave us to further investigate in future studies how strongly the observed effects influence the readability of tweets, and thereby build a useful model for filtering Twitter content for language learners.

References

Kevyn Collins-Thompson and James P Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 193–200. Association for Computational Linguistics.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability: Instructions. *Educational Research Bulletin*, 27(2):37–54.

James R. A. Davenport and Robert DeLine. 2014. The readability of tweets and their geographic correlation with education. *Computing Research Repository*, abs/1401.6058.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Thomas François and Cedrick Fairon. 2012. An AI readability formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 466–477. Association for Computational Linguistics.

Susan Homan, Margaret Hewitt, and Jean Linder. 1994. The development and validation of a formula for measuring single-sentence test item readability. *Journal of Educational Measuremen*, 31(4):349–358.

Robin Jeffries and Heather Desurvire. 1992. Usability testing vs. heuristic evaluation: was there a contest? *ACM SIGCHI Bulletin*, 24(4):39–41.

J. Peter Kincaid, Robert P. Fishburne Jr., Richard L.Rogers, and Brad S. Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Naval Technical Training Command.

Sigrid Klerke, Alexandra L. Uitdenbogerd, Falk Scholer, and Tim Baldwin. 2016. Twitter corpus from eye gaze study. Twitter data-set from an unpublished paper.

Marten Risius and Theresia Pape. 2016. Developing and evaluating a readability measure for microblogging communication. In *E-Life: Web-Enabled Convergence of Commerce, Work, and Social Life*, pages 217–221. Springer International Publishing.

Irina Temnikova, Sarah Vieweg, and Carlos Castillo. 2015. The case for readability of crisis communications in social media. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 1245–1250. Association for Computing Machinery.

Alexandra L. Uitdenbogerd. 2005. Readability of French as a foreign language and its uses. In *ADCS 2005: Proceedings of the Tenth Australasian Document Computing Symposium*, pages 19–25.

Sowmya Vajjala and Ivana Lucic. 2019. On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 349–359, Florence, Italy. Association for Computational Linguistics.

Carmen R. Wilson VanVoorhis and Betsy L. Morgan. 2007. Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, pages 43–50.