Register variation across text lengths
Evidence from social media

Aatu Liimatta
University of Helsinki

This paper explores variation in lexico-grammatical register features across text lengths in a large-scale sample of Reddit comments. Very short texts are known to be problematic for many statistical methods, so understanding their nature is important for the corpus-linguistic study of social media, where most contributions are short. I show that the frequencies of linguistic features change with comment length, even between longer comments, although longer texts are often considered similar in statistical terms. Moreover, I classify the variation found between short comments of different lengths into two main patterns, although other patterns can also be found, and there is variation even within these patterns. Furthermore, I interpret the observed differences in terms of register variation. For example, shorter comments appear to be more casual and less edited in terms of their feature makeup, whereas narrative and informational registers seem to favor longer comments.

Keywords: text length, register analysis, social media, Reddit, functional variation

1. Introduction

Short texts are widely recognized as problematic for corpus-linguistic studies which rely on per-word frequencies (also called ‘exposure rates’; Wallis, 2020: 47). Because of their shortness, even a single occurrence of a feature makes up a large fraction of the entire word count of the text. For example, a five-word text containing a single first-person pronoun can be calculated to contain 200 first-person pronouns per 1,000 words. There is arguably nothing unusual about a single first-person pronoun in a five-word text, but it is much more difficult to imagine an actual 1,000-word text with 200 first-person pronouns. These two cases are linguistically very different but have the same calculated feature frequency. Because
of this effect, frequencies like this are not very useful for many corpus-linguistic studies which include short texts.

Since normalized frequencies are nonetheless the dominant approach in corpus linguistics (Wallis, 2020: 47), researchers and compilers of corpora need to decide what to do about short texts in their data. This is particularly important for the corpus-linguistic study of social media, as most social media contributions are extremely short. There are several common ways of approaching the issue (as will be discussed in Section 2). However, in order to be able to make informed decisions about how to deal with short texts, and to understand what the effect of the various approaches may be, one needs to know what to expect from short texts to begin with.

In the present study, I will shed light on the nature of short texts by exploring the following two questions: (i) Are short texts and long texts fundamentally similar in linguistic terms, but this similarity is simply hidden due to the inherent difficulties in comparing them mathematically by means of e.g. normalization? (ii) Or are texts of different lengths different in some ways, and if so, in what ways?

If short and long texts are fundamentally similar, it would mean that shorter texts could be simply ignored even when they make up a majority of the analyzed dataset and the analysis could focus on the longer texts in the data. But if short and long texts prove to be linguistically different, the present study aims to serve as a step towards a better understanding of these differences, and towards more informed ways of approaching short texts in corpus-linguistic studies.

In the present study, I will look into the nature of very short texts, and the relationship between very short texts and longer texts, by comparing the frequencies of various linguistic features across comment lengths in a big data sample of social media comments. The background and the methodological approach to these questions will be explained in more detail in the following sections.

2. Register, social media, and text length

In this section, I first describe the multi-dimensional method of register analysis, which forms the basis this study builds on. Then, I explain the main problems caused by text length in corpus-linguistic studies, and how they are particularly relevant to the analysis of social media. Finally, I cover some earlier research on the effects of text length.
2.1 Register analysis

In order to explore the questions asked above and, more generally, the differences between texts of different lengths, the present study draws on the multi-dimensional method of register analysis (Biber, 1988). The method is based on the idea that differences in the requirements and circumstances of the situation and the purpose of the language use are reflected in variation of the frequencies of various lexico-grammatical linguistic features. In this model, text categories defined by their situational characteristics are called ‘registers’ (Biber, 1994; Biber & Conrad, 2001, 2009: 6). Registers can be defined at every level of generality (Biber & Conrad, 2001), and consequently register variation has been found for instance between large-scale high-level register categories across speech and writing (Biber, 1988), between historical subregisters in the news genre (Biber & Gray, 2013), on the internet more widely (e.g. Biber & Egbert, 2016, 2018; Titak & Roberson, 2013) and specifically between e.g. blog texts (Grieve et al., 2011), and even within individual social media platforms (Liimatta, 2019); the wide applicability of the multi-dimensional approach to different areas of language use is further illustrated by edited volumes, special issues, and meta-articles on the topic (e.g. Berber Sardinha & Veirano Pinto, 2014; Conrad & Biber, 2001; Friginal, 2013).

In terms of the present study, if texts of different lengths are different in their situation and purpose, one could expect to see variation in the frequencies of various register features across text lengths, reflecting the different production circumstances behind texts of different lengths. On the other hand, if feature frequencies are mostly stable across text lengths, it would suggest that text length is not an important variable in terms of register variation. Biber (2014) proposes ‘register universals’, dimensions of register variation found in nearly all multi-dimensional studies of register variation, no matter the dataset studied: a dimension of ‘oral’ vs. ‘literate’ discourse, and a dimension of ‘narrative’ vs. ‘non-narrative’ discourse. If any of the variation by text length found in the present study aligns with these universals, it further supports the link between variation by text length and register.

The focus of the present study is on register, although it is not a prototypical register analysis, in particular because text length has not been typically used to distinguish registers. While a more general label such as ‘functional variation’ could also have been used, I call the focus of the present study ‘register variation’. According to Biber and Conrad (2009: 2), “[a]ny text sample of any type can be analyzed from a register perspective”; furthermore, as would also be done in a more typical register study, the results of the present study are interpreted in terms of situational and functional differences between different text groupings, albeit
ones defined primarily by text length, and whose situational characteristics are explored a posteriori.

2.2 Social media and the problem of text length

Within the past few decades, information technology and computer-mediated communication have revolutionized communication through e.g. text messages, e-mail, and, more recently, the meteoric rise of social media which, together with the internet as a whole, provide corpus linguists with unprecedented access to large amounts of textual material. At the same time, the increasing interest in corpus-linguistic study of social media has brought the issues caused by varying text length and particularly short texts to the forefront for many working with such data, as most social media contributions tend to be extremely short. For example, within the data from the social media platform Reddit used in the present study, 50% of comments are under 16 words long, and 90% of comments are under 72 words long. On Twitter, before the tweet length limit was increased from 140 characters to 280 characters in late 2017 (see Eberl, 2020), most tweets were under 30 words long (Clarke & Grieve, 2017: 2) and probably not much longer, on average, after the length increase (see Rosen, 2017), though larger changes have taken place in some more focused Twitter datasets (e.g. Clarke & Grieve, 2019: 3).

When linguistic analysis uses methods which are subject to the confounding effects of short text length, there are two main solutions which are commonly used to deal with the problem. The first one is to simply remove from the dataset all texts shorter than some threshold, such as 400 words (e.g. Liimatta, 2020) or 1,000 words (e.g. Hiltunen, 2014). This is seen as a reasonable approach for datasets which contain mostly longer texts, since excluding the small fraction of short texts is considered unlikely to substantially affect the statistical linguistic picture of the dataset as a whole. However, this solution is clearly unsuitable for social media and other similar data, as an overwhelming majority of social media postings are below any commonly used text length threshold and would consequently need to be removed.

The second solution is to adopt a different operationalization of a ‘text’. For example, while Reddit comments can be considered individual texts, as is done in the present study, it is also possible to consider entire Reddit threads texts in themselves, polylogues with multiple participants (see Liimatta, 2019). This solution is reasonable when a higher textual level exists and is meaningful in terms of the research questions. Sometimes texts are also combined into larger units based on principled text groupings such as authors, text sources, registers, genres, or sociolinguistic groups. However, in order to use such groupings, one needs to have the metadata and a priori hypotheses to base the groupings on. Combining
texts in this manner will also obscure text-level patterns (Clarke & Grieve, 2017), so it is also necessary to consider the potential statistical implications. Egbert and Schnur (2018:160) advise against using “results from an entire corpus” (applicable in this case to a combined group of texts) instead of results from single texts, since such an approach may overstate the importance of features which are highly frequent in only a small fraction of the group and does not give much insight about generalizable patterns across the texts.

As the interest in the linguistics of social media has been increasing, methods which work with short texts are also being developed. For example, Clarke and Grieve (2017, 2019) use multiple correspondence analysis (MCA) to extract dimensions of variation of lexicogrammatical features in Twitter tweets. MCA is a technique specifically for analyzing co-occurrence patterns of categorical variables (see Glynn, 2014). While lexicogrammatical features are typically operationalized as continuous frequencies, they can also be made categorical by simply recording their presence or absence in a text. This works well with short texts such as tweets: because of their short length, it is unlikely that most features would appear in most such texts (Clarke & Grieve, 2017, 2019). At the same time, the method may not work that well with longer texts when studying high-frequency linguistic features such as lexicogrammatical features, as the longer a text is, the more chances there are for each feature to appear, theoretically eventually saturating the co-occurrence patterns when almost every feature co-occurs with almost every other feature. Although this effect has not been evaluated empirically in linguistic studies, even in their short-text Twitter data, Clarke and Grieve (2017, 2019) find text length to be a confounding factor.

Taken at face value, all discarding and combining approaches imply the view that short texts are, or should be, in every way equivalent to the longer texts within their reference group; that short texts are simply statistically problematic because of their shortness, but are actually fundamentally similar to longer texts, and so can be discarded or combined as one pleases without affecting the view of the underlying linguistic reality too much. Even when it is recognized that these options are a less-than-perfect necessary evil, one must work under the assumption that because short texts only make up a small fraction of the full data, whatever is done to them will only have an equally small effect on the results of the analysis. While this kind of mathematical and statistical reasoning can give us insight into the effects of such choices, how true this is, however, is in the end an empirical question; one which I believe is not possible to answer without knowing what short texts are like and what to expect from them linguistically to begin with. Furthermore, the validity of these assumptions is even more in question for social media and other genres made up of mainly short texts.
In order to take a step towards answering these fundamental questions about the effect of text length, the present study focuses on data from the social media platform Reddit. Reddit is made up of thousands upon thousands of “subreddits”, subforums dedicated to practically every imaginable topic, which are created, run, and moderated by Reddit users themselves. Reddit users can make posts in different subreddits, comment on the posts, and reply to other comments, creating tree-like comment threads under each post. The issue of text length is particularly important for Reddit data: the length of Reddit comments is not artificially limited, and consequently they are completely free to vary in length, leading to a range of comments of various lengths.

2.3 Earlier studies with text length as a variable

Even though, as described above, differing text length is recognized as a confounding variable in terms of the quantitative study of language, perhaps surprisingly, the effects of text length on linguistic variables have not been studied in much detail. Still, a number of studies related to the topic have been conducted. For instance, some studies consider the optimal length of text samples, particularly for the purpose of corpus construction. Notably, Biber (1993) looks at samples taken from longer texts in 200-word increments and shows, among other things, that common linguistic features differentiate genres well even with shorter samples, but rarer features require longer samples. According to Wallis (2020: 92), generally speaking a corpus with “many short texts is usually preferable to one with a few long texts”.

However, most studies on the effects of text length have focused on various measures of lexical diversity, most importantly the type-token ratio. This is because the type-token ratio is extremely strongly affected by the length of the text: due to the Zipfian nature of language, every additional word is considerably more likely to be one which already appears in the text, thus lowering the type-token ratio for longer texts and making it higher for shorter texts simply due to the mathematics of the situation. Hess et al. (1986) and Hess et al. (1989) demonstrate this strong confounding effect and show that there is no simple mathematical transformation which can remove it. While the type-token ratio is a simple and commonly used measure, because of these issues, various alternate measures of lexical diversity have been proposed. Koizumi and In'nami (2012) compare six different lexical diversity measures in short text samples between 50 and 200 tokens. They find that the best performing measure starts working with samples of 100 tokens and longer; thus, even the best measure in their study is unsuitable for most social media comments.
Even so, the nature of lexical diversity measures as a linguistic feature is very different from the more commonly studied lexico-grammatical features, and sample length as an explanatory variable is only tangentially related to text length. I am not aware of any study explicitly looking at the relationship between full text length and the frequencies of various linguistic features, particularly when it comes to shorter texts. Most published corpora have only included texts or text samples which are reasonably long; there have been no corpora or datasets large enough with a range of different text lengths to enable the study of variation by text length. I created a corpus of Reddit comments for this purpose, as detailed below in Section 3.1.

3. Data and method

This section describes the dataset and analysis methods used. First, I explain the source of the data, the data cleaning process, and the method used for determining the length of the text. Section 3.2 covers the methodology, which makes use of the large-scale dataset to circumvent many of the problems caused by text length.

3.1 Data

The present study makes use of a large-scale dataset of comments from the social media platform Reddit. The data originates from the so-called Reddit Comment Corpus, a constantly updated repository of all publicly available Reddit comments (Baumgartner et al., 2020). As the full Reddit Comment Corpus is extremely large in size and constantly growing, the present study is limited to a subset of the comments, viz. all publicly available comments posted on Reddit in August 2017. This small fraction of the entire Reddit Comment Corpus alone contains 84,658,503 comments.

The dataset required some cleaning to prepare it for tagging. First, comments including only the string “[deleted]” or “[removed]” were excluded, as these represent comments which had been removed from Reddit before they were collected for the Reddit Comment Corpus. The remaining comments were cleaned from various markup used to style the text for bolding, italics, quotations, internet links, etc. Furthermore, all characters which fall outside of Unicode’s Basic Multilingual Plane were removed, as they were found to crash the tagger; in practice, this means mainly emoji.¹ Finally, any comment left empty by the cleaning

¹ As also noted by an anonymous reviewer, these features (e.g. bold, italics, quotations, links, and emojis) would be important to include in a full MDA register study of similar data.
process, e.g. comments which only included a link or emoji, were also removed from the dataset. After this process, the size of the dataset had been reduced to 78,397,557 comments. The comments were then tagged for part of speech with the *Stanford CoreNLP pipeline tagger* (Manning et al., 2014) using the “english-caseless-left3words-distsim” tagging model.

Because the present study focuses specifically on text length as a variable, the method of calculating the length of a text is of particular importance. The most commonly used value for text length in corpus linguistics is word count or token count. These two terms are sometimes used interchangeably, but they can, and possibly should, be differentiated from each other. The simpler of the two to calculate is token count. The token count of a text can be acquired easily by simply counting the number of tokens the text has been split into by the tokenizer. However, text length acquired by such means is affected by e.g. the writer’s use of punctuation, as typically every punctuation character also counts as a token. The effect is especially strong with short texts: for example, in the case of a four-word text, the mere inclusion of a terminal punctuation mark increases the token count by 25%; or, conversely, its omission decreases the token count by 20%. Such a large relative change would also cause large differences in all derived values, including feature frequencies.

As the token count is a relatively volatile measure, it may be better to calculate a ‘word count’, which only includes tokens which would generally be considered ‘words’. For the purposes of the present study, the following types of tokens, as tokenized by the *Stanford CoreNLP pipeline tokenizer*, were counted as words:

i. Every token made up entirely of characters classified as Unicode letters, hyphens, and forward slashes. This rule includes typical words like *house* and *[high-grade]*, as well as words marked as interchangeable such as *big/large*.

ii. Every token made up either entirely of digits or of a mix of digits and letters, and possibly beginning with an apostrophe. This rule includes cardinal and ordinal numbers written in digits, such as *[2017]* or *[17th]*, as well as year and decade designations such as *[’98]* and *[’20s]*.

iii. The token *[’s]* when it is tagged as the verb *to be* in third person. In other words, this includes cases such as *she’s my friend* but excludes cases of *[’s]* as a possessive suffix such as *my friend’s house*.

iv. Every token which consists of an apostrophe followed by one or more letters between “a” and “z”, but which is not *[’s]*. This rule includes contracted forms such as *[’m]* for *am* in the count.

v. The token *[n’t]* for the contracted forms of *not*, such as *isn’t*. 
The relatively simple set of rules listed above covers most major cases. It is of course not always straightforward to decide which tokens exactly should be counted as words and which not (see e.g. Baroni, 2008). For example, the number of tokens, and so the number of words counted by the above rules, may change depending on hyphenation. There are also arguments to be made for counting some multi-token units as single words or counting certain single tokens as multiple words.

The rules used in the present study are based on the principle that every token either is or is not counted as a single word, and the rules specify which tokens exactly to count as words. I believe these rules, while by no means perfect, help form a more principled baseline for word count than simply including every token in the calculation. My goal in creating these rules, after all, is not to create a universal definition of a ‘word’, but rather to include in the analysis most of the kinds of tokens which would generally be considered words and exclude most of the kinds of tokens which would generally not, in order to reduce noise in the data for the computational analysis. Furthermore, it is useful to know what exactly the word counts in a study are based on for both replicability purposes as well as increased understanding of what exactly the data is telling us, even if, or particularly when, it is often not stated very clearly.

3.2 Method

In order to determine the influence that comment length has on feature frequencies, the present study utilizes a method building on so-called ‘lengthwise approaches’ (see Liimatta, 2020). The idea behind this family of methods is that while texts which are different in length can be difficult to compare with each other (as discussed above), texts which are the exact same length can be compared trivially. In other words, every lengthwise study is made up of two separate steps. In the first step, a suitable method chosen by the researcher is applied to texts of the same length (‘intra-length’), for each length separately (‘lengthwise’). After this, in the second step, the results can be compared across text lengths (‘inter-length’) using some suitable method.

The present study uses a simple but powerful lengthwise pooling method. For each of the features studied, the average frequency of the feature was calculated for each comment length separately by dividing the total number of occurrences of the feature in comments of a specific length by the total number of words within comments of that length. For example, for the length of 150 words, the dataset contains 29,347 comments, which contain a total of 81,673 infinitive forms; thus, the frequency of infinitives in 150-word comments is $\frac{81,637}{(29,374 \times 150)} = 0.0185$. 

Aatu Liimatta
This kind of analysis allows us to look at the larger picture of comments of every specific length, avoiding the typical mathematical problems caused by very short texts. In a sense, this method ignores feature frequencies in singular comments, with all of the related problems, to get a view of the large-scale tendencies across comment lengths. It is then possible to create figures which show the change in the average frequency of the features as a function of comment length.

This approach requires a large dataset, as there need to be enough texts of every single length to draw meaningful conclusions from (see also Liimatta, 2020). Large social media datasets, such as the one-month Reddit dataset used in the present study, are ideal for this. Even so, some binning of longer comments was necessary even with such a large dataset in order to reduce noise caused by the low number of samples for longer comment lengths. In practice, bins were created by starting from the shortest comments, and combining comments of increasing lengths so that every bin contains at least 1,000 comments. For lengths up to 426 words, this means that every bin only consists of comments of the exact same length, with thousands, tens of thousands, or even millions of comments within one bin; whereas with longer texts, adjacent comment lengths were typically combined into one bin. For instance, comments with 429 and 430 words were combined, as there are only 967 and 970 of them, respectively; the combined bin of 1,937 comments exceeds the minimum requirement of 1,000 comments. While binning may be at odds with the idea of lengthwise analysis, it is unlikely to affect the overall picture for longer comments very much, as an absolute difference of a few words between longer texts is only a small difference in relative terms: it is still possible to see the overall trends for longer comments from the results. However, binning very short texts across text lengths should of course be avoided, as even a single-word difference represents a large fraction of the full length of a short text.

For the present study, 61 features based on those used by Biber (1988) were investigated. The number of occurrences of each of the features was counted for every comment in the dataset mainly using the recognition algorithms given in Biber (1988). The tagger used provided more detailed information than needed for the Biber (1988) algorithms, enabling easier recognition of some features. The following features difficult to detect using fully automated means were excluded: gerunds, present participial clauses, past participial clauses, present participial WHIZ deletion relatives, and sentence relatives. Furthermore, as the focus of the present study is on feature frequencies, type-token ratio and mean word length were also excluded. Finally, first-person pronouns was split into first-person singular pronouns and first-person plural pronouns, as they had exhibited different behaviors in earlier trials.
In the following section, I will show examples of specific features which exhibit different kinds of variation between comments of different lengths and give functional interpretations for these observed differences.

4. Variation across text lengths

Figure 1 is a zoomed-in view of the average frequencies of all of the investigated features across Reddit comments between 1 and about 200 words in length. The vertical lines show the median and 90% quantile comment lengths, respectively. Figure 1 is a high-level overview of the results; the observed patterns are discussed in more detail in the following subsections. The features have been colored to make them easier to distinguish from each other. A handful of the features have frequencies far above most, and as such are cut off in the figure to focus on the majority of the features.

The spikes at specific points on the feature frequency curves are often a result of automated bot comments and repeated comment templates used by subreddit moderators. Such comments are used for many reasons, such as to remind commenters about subreddit rules, or to tell the reason a rule-breaking submission was removed. As the exact same, or very similar, comments get posted repeatedly, the feature frequencies for certain comment lengths get skewed by the feature makeup of these posts. However, in order to analyze the differences between comments of different lengths in general, one can simply ignore the spikes and follow the overall shape of the curves.

Figure 1 shows that in the comments at the longer end of this length range, the average frequencies stay relatively stable regardless of comment length. However, the picture is entirely different for the shorter comments, which show large changes in feature frequencies even with relatively small absolute differences in comment length. Furthermore, not every feature follows a similar trajectory, as might be expected if these patterns were simply a mathematical artifact.

The features appear to follow two main patterns, labeled here the ‘rounded rise’ pattern and the ‘initial peak’ pattern, which are discussed in more detail below. It bears repeating that while these differences are observed between extremely short comments, these comment lengths are extremely common on Reddit. In other words, while the changes cover only a relatively small portion of the length range shown in the figures, they actually apply to a vast majority of the comments in the dataset, as also shown by the vertical lines indicating the median and 90% quantile comment lengths.

In the following subsections, I focus on the variation in a selection of the features. These features were selected as representative examples of the kinds of vari-
ation found within the data in general, and also to illustrate the connection of the variation between comment lengths with register variation. Furthermore, I include excerpts from the Reddit data to illustrate the use of the features on Reddit and to give a better concrete understanding of the types of registers I talk about.

4.1 Rounded rise

The most common pattern within the data is the ‘rounded rise’, accounting for roughly 40% of all of the features studied. Visually, this pattern is characterized by a very low, close-to-zero frequency of the feature in the very shortest comments, which rises quickly but with a smooth curve to roughly match the baseline frequency found in the longer comments. In other words, the features following this pattern are rare in the very shortest comments and increase in frequency as the comments get slightly longer.

Figures 2 and 3 show the frequencies of infinitives and third person pronouns, respectively. Both features follow the rounded rise pattern, with an initial rapid rise which then smoothly slows down to meet the baseline frequency, the frequency of the feature in longer comments, which stays relatively stable across comment lengths, at least compared to the large changes which take place within the shortest comments (but see Section 4.4).

Of the patterns explored in the present study, the rounded rise pattern is perhaps the one most easily explained by the simple mathematical constraints of the linguistic situation, viz. that a text needs to be long enough for it to have enough
Figure 2. Frequency of infinitives across comment lengths from 1 to 200 words

Figure 3. Frequency of third person pronouns across comment lengths from 1 to 200 words

room for any specific feature to appear. In other words, given any extremely short comment of maybe under 10 words or so, it is quite unlikely that any of its words
will be a third person pronoun, or that it will contain infinitives, as there is nei-
ther space nor the need for such features in such short texts. But the longer the
comment, the higher the probability of the presence of any such feature becomes,
to the point that when the comment gets long enough, the feature starts appearing
at a certain frequency, which levels off the graph.

This pattern is also illustrated by the split auxiliaries in Figure 4. While the feature
is relatively rare overall, occurring only under 2.5 times per 1,000 words in the
data, it too follows the rounded rise pattern, as the comment simply needs to be
long enough so that it even can include an auxiliary construction which can be
split.

Examples (1) and (2) show how infinitives are used in the data. All examples
only include full Reddit comments, not excerpts from longer comments. Bolding
is used to highlight the features, and quotation marks delineate comments from
each other. In Example (1), the original poster shows an empty package they
received. In Example (2), the first comment is answering a question about suppos-
edly nice gestures which one nonetheless finds annoying; the reply to this points
out that sorry also has other meanings than to apologize.

(1) “Lmao was there nothing in there?”
   “Just a packing slip.”
   “I want to know how customer service handles this one”
“People apologizing for things that aren’t their fault because it seems like the nice thing to do.”

“Sorry isn’t just to apologize, it’s also used to sympathize with someone”

Examples (3) and (4) show two different uses of third person pronouns referring to people outside of the immediate context of the conversation. In Example (3), in answers to a question about your own annoying habits, third person pronouns are used to refer to people outside of Reddit, to other people in general in the first comment, and to a specific person, the commenter’s wife, in the reply to it. In Example (4), third person pronouns are used to refer to another Reddit user in the conversation. The original post was a request for other users to check the parts list the poster has made for a new PC they will be building. The first commenter suggests changes to the list. The reply to this argues that the suggested parts are not compatible, referring to the original poster using the third person.

(3) “Cutting people off mid sentence because I’m so excited to tell them something that can add to the topic...”

“This. It drives my wife absolutely insane and she thinks I’m being really rude.”

(4) “Here. Ive replaced the ram and saved you $15, swapped the case to what i think is much nicer. You can swap it back if you like. And That PSU is a really good deal.”

“That ram isn’t compatible with his mobo”

Example (5) shows a common use of a split auxiliary strengthening the force of a claim.

(5) “This has been posted countless times in the past hour”

“I honestly didn’t even see it sorry”

In addition to the infinitives, third person pronouns, and split auxiliaries discussed above, some other features following the rounded rise pattern include downtoners, existential there, indefinite pronouns, perfect aspect, possibility modals, and predictive modals.

4.2 Initial peak

The ‘initial peak’ pattern accounts for roughly 1/6 of the features studied. While the feature is less common than the rounded rise, it is much easier to recognize and differentiate from other patterns. Visually, this pattern is characterized by a sharp peak at the very shortest end of the length range, after which the frequency quickly falls. In other words, the features which follow this pattern have a
markedly high frequency in very short comments, but the frequency sharply falls to the baseline level as the comments get longer.

Figure 5. Frequency of direct WH-questions across comment lengths from 1 to 200 words

Figure 6. Frequency of second person pronouns across comment lengths from 1 to 200 words
Figures 5 and 6 show the frequencies of direct WH-questions and second person pronouns, respectively, across comment lengths within the data. In contrast with the features in the previous section, which follow the rounded rise pattern, both of these features exhibit an obvious peak in the shortest comments. This jump to a high peak frequency followed by a relatively sharp drop to a lower baseline frequency contrasts with the rounded rise pattern, which has a relatively smooth rise to a higher baselevel frequency.

These differences can be seen to reflect register differences between comments of different lengths. For example, a high frequency of direct WH-questions in short comments (Figure 5) can be explained through functional means. A simple information-seeking question does not usually have to be particularly long, as shown in Examples (6) and (7). At the same time, longer questions have a lower frequency of direct WH-questions, since in addition to a simple request for information, the comment is likely to include other background information, such as an elaboration of the reason for asking the question, an explanation of the question asker’s position on the topic, or other such content not directly information-seeking in nature. However, in such cases, the function and other situational characteristics of the comment can be understood to be slightly different from a comment with only a short information-seeking question. The short comment is thus the natural mode for the direct to-the-point information-seeking question register, and as such it is natural that the shortest comments also contain the highest frequency of direct WH-questions.

(6) “Definitely pork scratchings. They are amazing.”
“What is a pork scratching?”

(7) “I wish I could get beer at a gas station..”
“Where are you that you can’t?”

The pattern of the second person pronouns in Figure 6 can also be explained in a similar manner. Pronouns in general mark a lower density of information, less formal style, and an interpersonal focus; second person pronouns in particular indicate “a high degree of involvement” with a specific addressee (Biber, 1988: 225). In the typical case on Reddit, second person pronouns are used to refer to the writer of the comment being replied to, or sometimes a generic you. Consequently, the baseline frequency of second person pronouns is already relatively high, as many longer comments also use the generic you or speak directly to the previous commenter. What further elevates the frequency of second person pronouns in the shortest comments is the tendency to reply to something the previous commenter has said with only a short reaction comment using the second person pronoun. Such usage comes on top of the baseline level of second person pronoun use. Furthermore, second person pronouns tie together with WH-
questions, as the short reaction comment may also be asking further questions about a personal topic the previous comment talked about.

Example (4) above uses the second person pronoun for interpersonal references to the user who posted the original parts list, and Example (7) targets the question to the previous commenter using the second person pronoun. Example (8) is a part of a discussion about hobbies which one would like to pursue full-time if financial security was not a problem. The first and third comments are by the same person; the other three comments are all by different users. Second person pronouns are used in interpersonal functions to refer to the previous commenter. Finally, Example (9) shows two examples of the generic you.

(8) “Photography, such an amazing art form.”
    “Why aren’t you doing it, then?”
    “I just don’t have time. I work all the time.”
    “if you wanted to do it you would do it. You would just rather work all the time.”
    “Gee, you really got it all figured out don’t you”

In Example (9), both comments use the generic you, first as a part of a motivational quote, and then in a paraphrase of its meaning showing agreement with the previous commenter.

(9) “Also, “time you enjoy wasting isn’t wasted time.””
    “That is also true. If you enjoyed the time it was a good investment.”

Figure 7. Frequency of amplifiers across comment lengths from 1 to 200 words
Similarly, the frequency of amplifiers in Figure 7 shows a peaked pattern. Amplifiers typically strengthen the force of verbs, adjectives, and adverbs, and they are used to indicate the reliability of a proposition and to signal solidarity with the addressee (Biber, 1988: 240). Like second person pronouns, they are more common in the shortest comments, which are often quick reactions to something said earlier, but they also see some use in longer comments. Examples (10) through (12) show how amplifiers are used to strengthen statements on Reddit. In Example (10), an amplifier strengthens the commenter’s agreement with the previous commenter’s evaluation of a character in a television show. Example (11) is very prototypical, amplifying a positive evaluation. In Example (12), in a discussion about Karl Marx’s grave, the commenter amplifies the strength of their proposition.

(10) “He makes a lot of good faces this episode.”
   “I totally agree”

(11) “I haven’t been drawing in ages, but this inspired me to make a quick sketch.”
   “Wow! Very nice! thank you!”

(12) “Highgate cemetery is a place for many famous people.
   Some I can remember – Karl Marx, George Michael, Alexander Litvinenko,
   Michael Faraday, George Eliot, Douglas Adams”
   “I didn’t realise any of the others, may have to visit those too!”
   “It’s definitely worth it. Highly recommended.”

As seen above, the exact shape of the initial peak pattern varies from feature to feature, depending on how exactly the function of the feature affects its use in shorter and longer comments. In fact, the pattern can be divided into at least three subtypes. Firstly, in some cases, the baselevel frequency is at a near-zero level, indicating that the feature is almost nonexistent within longer comments, and only appears at any considerable frequency within the very shortest comments. The direct WH-questions (Figure 5) follow this pattern. Secondly, for some other features, the baselevel frequency is some low but non-negligible frequency: the feature is much more frequent in the shortest comments but does appear at some frequency even within longer comments. Of the feature examples above, amplifiers (Figure 7) are the closest to this pattern. In both of the above cases, the peak frequency of the feature is at least double but typically many times more than the baseline frequency. Thirdly, in yet other cases, the baseline frequency itself is

---

2. In the present study, in accordance with Biber (1988: 240), the following items were counted as amplifiers: absolutely, altogether, completely, enormously, entirely, extremely, fully, greatly, highly, intensely, perfectly, strongly, thoroughly, totally, utterly, very.
already relatively high, and the peak frequency only rises above the baseline frequency by some smaller but still substantial amount; in such cases, the feature is also commonly used in longer comments, but is simply more common in the very shortest comments. This is the case with the second person pronouns (Figure 6). In addition to the direct WH-questions, second person pronouns, and amplifiers discussed above, some other features following the initial peak pattern include \textit{be} as main verb, demonstrative pronouns, and first person singular pronouns.

4.3 Other patterns

Any classification of such patterns of course depends on the desired granularity of the classification system, and as there is considerable variation between individual features, is bound to have fuzzy boundaries. However, some features are quite distinct in the individual characteristics of their frequency distribution across comment lengths. In this section, I will show two features whose patterns do not match either of the two main patterns very well.

![Figure 8. Frequency of nominalizations across comment lengths from 1 to 200 words](image)

Figure 8 shows that the frequency of nominalizations increases gradually as the comment length increases, reaching its high baselevel frequency at around the 90% quantile of comment length. Nominalizations are used to integrate information into fewer words (Biber, 1988: 227), and are associated with a literate, as opposed to a spoken, style across various genres (Biber, 2014). This has also been
shown to apply to Reddit (Liimatta, 2019). For the purposes of the automated feature recognition, nominalizations are defined as words ending in -tion, -ment, -ness, or -ity, or their plural forms, in accordance with Biber (1988: 227), and further refined to only words which are at least two characters longer than the suffix. This approach makes it simple to recognize the feature programmatically; however, it does not take into account other types of nominalization, such as zero derivation, and some words are mistakenly recognized as nominalizations. Nonetheless, the aim is to optimize for the precision of recognition while sacrificing recall. While we do not get the actual frequency of all nominalizations this way, the results are still comparable for approximating differences in the use of nominalizations across text lengths.

The correlation between the frequency of nominalizations and the comment length is thus naturally linked to register differences between comments of different lengths. It is reasonable to think that short comments, where the frequency of nominalizations is lower, tend to be closer to spoken language, often just quick casual comments or reactions to something said earlier. Equally reasonable is the hypothesis that the longer a comment becomes, the more time and effort the writer expends to edit the comment to include all of the relevant information in the way they want to. This means that longer comments are likely to be more informational, more edited, and more “literate” than shorter comments.

Example (13) shows the use of nominalizations in a technical explanation on Reddit:

(13) “Why do these replays from the late 90s/early 2000s look like they were made in the 70s? I know we had better quality TV than this.”

“Resolution was TV 640 × 480 or worse, years of reproduction eventually leads to degradation (not from masters but reproductions – also time before they were put on digital loses some clarity), 4:3 aspect only, on screen graphics look ancient, etc.”

The frequency of contractions, shown in Figure 9, is also related to the literate-oral register dimension. The frequency of contractions starts at a relatively high level, but gradually decreases as the comments get longer. Contractions are commonly used in casual, colloquial comments, but as they are dispreferred in formal, edited writing (Biber, 1988: 243), their frequency is likely to decrease the more time and effort the writer spends on thinking about what and how they write, as is naturally the case with longer comments.

Examples (8), (11), and (12) above also illustrate some typical contractions found in casual shorter comments. In addition to nominalizations and contractions, features such as conditional adverbial subordinators if & unless, hedges, past tense and private verbs follow patterns which cannot be as easily classified.
Lengthwise variation between longer comments

Short texts are a particular focus of the present study, as they are extremely frequent on social media but less studied in previous work, and often a problem when applying statistical methods. Still, one must not forget about longer texts, as the effects of text length on longer texts are not fully known either. Longer texts are generally considered comparable with each other as long as their feature counts are normalized. However, the present analysis shows that, depending on the feature, this may not always be the case: while the frequencies of many features stabilize when the comments get longer, there are also many features whose frequencies change with comment length even between longer texts. To enable comparison between longer comments, Figures 10 through 13 expand the horizontal scale, showing the frequencies of third person pronouns, past tense forms, second person pronouns, and predicative adjectives, respectively, to 1,500 words.

For many features, the frequency stays relatively stable within longer comments. Predicative adjectives in Figure 13 are an example of this. After the initial peak in the very shortest comments, their frequency stays at a highly stable baseline frequency of around 0.005. But the frequencies of many other features change quite dramatically even between longer comments of different lengths, as illustrated in Figures 10 through 12.
**Figure 10.** Frequency of third person pronouns across comment lengths from 1 to 1,500 words

**Figure 11.** Frequency of past tense forms across comment lengths from 1 to 1,500 words
Figure 12. Frequency of second person pronouns across comment lengths from 1 to 1,500 words

Figure 13. Frequency of predicative adjectives across comment lengths from 1 to 1,500 words
In Figure 10, the frequency of third person pronouns increases by almost 50% between the very short comments and comments 1,500 words long. Similarly, the frequency of past tense forms also increases with longer comments, as shown in Figure 11. Both third person pronouns and past tense verb forms are common features of narrative registers (Biber, 2014), implying that narrative registers tend to favor longer comments relative to other registers on Reddit. This might be expected, as various types of narratives need longer stretches of text to tell the story, oftentimes a lot longer than the median comment length, and common narrative types such as fictional stories and factual descriptions of various events commonly use past tense verb forms and refer to the characters and people appearing in them using third person pronouns. The use of these features is illustrated in Example (17).

On the other hand, second person pronouns in Figure 12 show a different kind of development: their frequency decreases as the comments get longer. This is explained at least in part by the opposite of the above, viz. that second person pronouns tend not to be as common, in relative terms, within narrative contexts, but are more common in interpersonal and instructional contexts. Such registers might favor shorter comments on average, possibly reflecting typical spoken interaction as exchange of “alternating (and relatively short) bursts of information” (Holler et al., 2015:1).

Examples (14) and (15) show some uses of predicative adjectives. In Example (14), the commenter describes fast food. In Example (15), the second commenter is directed to a subreddit which would be better suited for their plant identification request; in response, they praise the other commenter, and give a reason for not being aware of a better subreddit.

(14) “I know fast food is horrible for you, but every now and then it just hits that comfort/nostalgic spot”

(15) “Post in r/whatsthisplant for proper I.D.”
   “You are awesome! I am newish to reddit. Well, just started again using reddit.”

Examples (16) and (17) show some uses of past tenses. In Example (16), the second commenter describes their personal past experience; in Example (17), discussing a photograph of the German politician Angela Merkel visiting fishermen when campaigning in 1990, the second commenter relates details about the meeting which they remember reading.

(16) “Reality tv is scripted”
   “Yeah I was obsessed with Repomen and then hated it when I found out the truth..”
“What was she saying to them? Is there any record of the conversation?”
“I don’t know if the conversation was recorded, but the news article I read about it (who talked with one of the fishermen) said that she downed alcohol with them and didn’t talk much about herself but asked more about them in the beginning. Later she started talking a bit about her political party, and at the goodbye she promised them she would try to do something for them / the fishing-industry. She left with a good enough impression that they voted for her, as far as I understood it.”

5. Conclusions

The present study shows differences in frequencies of a large number of lexico-grammatical features between Reddit comments of different lengths within the large-scale Reddit dataset analyzed, both within the very shortest comments and between shorter and longer comments. The lengthwise differences, i.e. differences between text lengths, observed between the shortest comments can be classified into two main patterns, which together cover over 50% of all features studied. The features following the rounded rise pattern have a very low frequency in the very shortest comments, but their frequency quickly but smoothly increases to meet the baseline frequency of the feature as the comments get longer. The initial peak pattern, on the other hand, has the highest frequency within the very shortest comments, with a sharp decline in frequency to meet the baseline frequency of the feature as the comments get longer. But there is also internal variation within these patterns, and other kinds of patterns can be found in addition to these two main patterns. Furthermore, variation in the frequencies of linguistic features can also be found across longer comments, where the frequencies typically have a generally increasing, decreasing, or stable trend as the comments get longer.

These observed patterns point to register differences between comments of different lengths. For example, the initial peak in direct WH-questions suggests that simple information-seeking questions favor short comments on Reddit, and the initial peak of second person pronouns shows that shorter comments tend to be on average more interpersonal in their focus. The increase of nominalizations and decrease in contractions as the comments get longer reflects that, on average, longer comments tend to be more informational in their focus and more edited in their style, and that shorter comments tend to be more casual, colloquial, and unedited. Similarly, the higher frequency of past tense forms and third person pronouns in longer comments suggests that narrative content tends to favor longer comment lengths. Both informational and narrative concerns favor longer com-
ments as longer length is a natural consequence of giving a detailed account of what happened or delivering detailed information.

Since different features follow different patterns and their frequencies change at different rates as the comment length changes, this also means that comments of different lengths have different mixes of features. Furthermore, many features have nonmonotonic patterns which reach a clear peak or bottom and change direction. This further increases the different relative mixes of register feature frequencies for different comment lengths. In other words, the effect of comment length is not a simple relationship where short comments have one mix of register features and longer comments have another mix. Instead, the relative frequencies of many of the features keep changing across comment lengths, hinting at many different registers which favor different comment lengths.

The results of the present study are also in line with the proposed register universals (Biber, 2014), viz. they exhibit a dimension of ‘oral’ vs. ‘literate’ discourse and a dimension of ‘narrative’ vs. ‘non-narrative’ discourse. However, unlike in earlier work, the present study finds this variation between comments of different lengths, confirming that the register universals also seem to apply to register variation observed across text lengths.

Even so, the exact relationship between feature frequency variation by text length and various other variables, including register, is still unclear and remains a subject for future studies. For example, do different registers (as they are typically understood, including texts of different lengths) also exhibit internal variation by text length, or do different registers or text types simply favor different text lengths, leading to apparent text-length-based variation in a big picture view? Text length is naturally linked to register variation, chosen by the writer to match the situational requirements. For example, in some situations, the writer can choose the length of the text relatively freely to suit their needs, whereas in other situations, the length is more or less limited by e.g. the physical medium (like in the case of postcards or letters), genre conventions or publisher requirements (e.g. classified advertisements or newspaper articles). After all, short texts and long texts are written for different purposes and in different situations, and consequently would be expected to exhibit register differences.

I would argue that text length has even more important implications for corpus-linguistic studies than has thus far been commonly recognized. A wealth of genres, including social media, are made up of largely short texts. In order to study such genres effectively with many quantitative approaches, particularly when the data being studied contains a mix of both short and longer texts, we need to develop methods which can get around the problems with applying more traditional quantitative corpus-linguistic methods to them. However, in order to be able to do that in an informed manner, one first needs to know what to even
expect from such texts. The present study represents a single step in that direction. In the meantime, the corpus linguist at least needs to be aware of the possibility of the existence of lengthwise variation within the data, and consider if and how to best account for it in their analysis.

Acknowledgements

I would like to thank everybody who read the manuscript, particularly the two anonymous reviewers, for their invaluable comments.

References


