Text length and short texts : An overview of the problem

Liimatta, Aatu

2024-09-19

Liimatta, A 2024, Text length and short texts : An overview of the problem. in M Kaunisto & M Schilk (eds), Challenges in Corpus Linguistics : Rethinking corpus compilation and analysis. Studies in Corpus Linguistics, vol. 118, John Benjamins, Amsterdam, pp. 106-125. https://doi.org/10.1075/s

http://hdl.handle.net/10138/601562
10.1075/scl.118.07lii

unspecified
acceptedVersion

Text length and short texts: An overview of the problem

Aatu Liimatta

University of Helsinki

Abstract

Variation in text length is an unavoidable confounder in quantitative text-analytic corpus-linguistic studies. Texts can be difficult to compare across text lengths, particularly if many of them are short, due to the difficulty of calculating meaningful frequencies for the lexical items and linguistic features of interest. Traditionally, this problem has been less of an issue, since texts in many of the genres typically studied in linguistics have been relatively long. However, the rise of social media has brought the issue to the forefront. In this chapter, I describe the problem of text length and short texts, and a number of solutions and workarounds to these and related problems. I also propose some potential improvements and new approaches.

1. Introduction

It is a well-known fact that variation in text length is an unavoidable source of nonuniformity in corpora and can cause issues in quantitative corpus-linguistic analyses. At the very basic level, the confounding effect of variation in text length is obvious. Since a longer text by definition contains more words, there are also more opportunities for any given linguistic item or feature to appear. In other words, a longer text will, on average, contain more instances of any item or feature simply because it is longer. This is a problem particularly for text-analytic corpus-linguistic studies, which are interested in comparing how many of these items appear

in different types of texts: if two texts have a different number of occurrences of the feature of interest simply because they are of different lengths, it can be very difficult to compare texts of different lengths with each other.[1]

Fortunately, this basic problem has a simple mathematical solution, which commonly forms the basis of typical quantitative corpus-linguistic inquiries. The number of occurrences of the feature of interest in a text can be divided by the number of words[2] in the text. This process, *normalization,* gives us the rate of occurrence of the feature, i.e. the average number of instances of the feature *per each word in the text*. Viewed differently, this normalized value can be understood as the probability of picking an occurrence of the item of interest if a word is picked from the text at random. Commonly, this value is further multiplied by a factor of e.g. a thousand, ten thousand, or a million, depending on the overall prevalence of the feature in the dataset, in order to bring it within a more understandable numerical range. Crucially, the normalization process harmonizes the divisors of the feature counts from texts of different lengths. For instance, 10 occurrences per 2,500 words and 20 occurrences per 3,200 words both get the same divisor after normalization, e.g. 4 occurrences per 1,000 words and 6.25 occurrences per 1,000 words, respectively. This makes it possible to directly compare the texts which these rates of occurrence represent.

Normalization is a tried-and-true method of comparing texts of different lengths with each other. However, while it is a working solution to a huge potential problem in a large number of cases, normalization is not without problems itself. One problem with normalization becomes particularly salient when applying the method to short texts. The problem is based on the mathematical fact that the smaller the divisor becomes, the larger the result of the division will

[1] Variationist corpus linguistics, which focuses on the proportions of variant items or constructions, is not as heavily affected by the issue. However, even variationist analyses may be affected by the distribution of text lengths in their dataset.

[2] Many other bases of comparison other than words can also be used. However, the focus of the present chapter is on word count specifically, since it is the most commonly used basis in most quantitative corpus-linguistic studies.

be. In other words, the fewer words there are in a text, the larger the normalized value is. This is of course the very basis on which the normalization method is built to enable comparisons of texts of different lengths. However, when the divisor becomes very small, the effect gets magnified and the result of the calculation inflates to meaningless levels. For instance, consider a short text of only five words (such as a tweet, a postcard, or a sticky note) which contains one instance of a feature, for example a single first person pronoun. If we calculate the normalized frequency of first-person pronouns in this text, we get as the result 200 first-person pronouns per 1,000 words. This is the mathematical solution to the formula, but the result is quite useless in terms of comparing texts of different lengths with each other. While not every five-word text contains a first-person pronoun, it is also not that unusual if one does. But the same normalized rate of occurrence value also applies to a text of 1,000 words which contains 200 first person pronouns. While both the five-word text and the 1,000 word text have the same rate of occurrence of first person pronouns, surely the 1,000 word text with 200 first person pronouns is much more unusual than the five-word text with one first person pronoun. Clearly, these calculated rates of occurrence are not meaningful measures for the comparison of short and longer texts in linguistics.

In other words, there are two related problems caused by text length. First, texts of different lengths cannot be directly compared because they have different numbers of everything simply due to their difference in length. I call this the *problem of text length*. Second, extremely short texts cannot be easily compared with other texts using many typical quantitative corpus-linguistic methods, such as normalization, because the normalization results become meaningless as the length of the text becomes increasingly short. I call this the *problem of short texts*. In order to talk about these problems, in this chapter, I use the words “long” and “short” to refer to texts which are and are not long enough for typical quantitative corpus-linguistic analysis, respectively, though the line between the two is of course fuzzy.

To combat these two problems, a number of solutions and workarounds of varying sophistication have been devised. However, many of these solutions have problems of their own. Despite the ubiquity of the problem, and the often-suboptimal nature of its solutions, the problem is not that often discussed in much depth. In this chapter, my goal is to bring more attention to the problems of text length and short texts, and to encourage the development and application of new and improved approaches to the problem. In Section 2, I describe how the problem of text length was historically less of an issue but is coming to the forefront with the rise of research into the language of social media. I also refer to and summarize results from earlier studies, which suggest that texts of all lengths are of interest. In Section 3, I cover various methods which have been used to either solve or work around the issues caused by the problem of short texts, the problem of text length, and related problems, and discuss their upsides and downsides, as well as suggest best practices and propose potential improvements to these methods. Furthermore, I will briefly discuss the related topic of the effect of text length on measures of lexical diversity, which has been studied in more detail. Finally, in Section 4, I will conclude this chapter with some final thoughts on the problems caused by text length and their solutions.

2. Background

2.1 Text length, corpora, and social media

The problem of text length and short texts is caused by a simple mathematical relationship, and as such it has been known of since the beginning of quantitative corpus linguistics. However, historically, the practical problems it has resulted in have arguably been of relatively little

actual consequence. Most genres traditionally studied in quantitative corpus linguistics tend to comprise of longer texts, compared to the extremely short text lengths of up to only a few dozen words which are overwhelmingly common on e.g. social media. Because of the comparatively long text length in such genres, the normalization method works reasonably well with them.

The reasons for the focus on genres with "longer" texts have been manifold. A major reason has of course been, and still is, availability of data. Researchers have to use whatever data they have access to. Historically, this has largely been published corpora compiled by teams of researchers. But the compilers of such corpora have also been working with the data they can get access to in large enough quantities. These include genres such as newspaper articles, fiction writing, academic papers, and countless others. Many of these genres have editorial guidelines or genre conventions which place certain requirements for the length of the piece of writing.

Another reason for the focus on longer texts is what has been considered important and influential enough to study. The impact of e.g. newspaper articles, academic writing, casual conversations and personal letters on people, society, and language has been evident to all, and as such it is only natural that texts in such genres are of interest to anyone studying language. But these texts also tend to be long enough for reasonable quantitative corpus-linguistic analysis. It is often easier to overlook the societal and linguistic impact of genres with mainly shorter texts. For instance, personal notes, post cards, and shopping lists are also something written and read regularly, but we might not even think to consider them and other similar genres as research subjects.

The question of influence and significance also comes back to the question of availability, since it is of course more difficult to collect a large and representative corpus of post cards or shopping lists than of newspaper articles. However, it is also, to an extent, a chicken-and-egg situation. With more interest in such genres, it is possible that more such data would be

collected into corpora; and with more such corpora available, there might be more interest in such genres.

A similar vicious circle has also developed when it comes to the development of analysis methods which would allow us to better approach genres with short texts. If genres with longer texts are easier to quantify and the available methods work better with them, it is only natural to focus on such genres; and if the focus typically is on genres with longer texts, there is no particular need to develop methods and approaches which could help analyze genres with shorter texts in more detail.

However, over the past decades, many of these paradigms have been gradually but firmly upended. A central catalyst for the change has been the spread of the internet. The web and other means of computer-mediated communication have become easily accessible sources of linguistic data, which has greatly facilitated the building of new corpora and datasets to match the needs of the researcher and to enable the study of entirely new kinds of genres and registers. Of course, published corpora are still being compiled to this day, and they are a valuable tool used in a wide variety of linguistic research. But the easy access to textual data online has allowed quantitative corpus linguists to cast a wider net in terms of their research topics than ever before.

At the same time, the rise of web and CMC texts has brought many of the issues with text length to the forefront. In contrast to the more "traditional" genres which make up many compiled corpora, texts from many internet genres tend to be less bound by word count limits or guidelines. While many online genres have a highly variable text length and a large proportion of shorter texts, such as blog posts or Wikipedia articles, this is particularly true for computer-mediated communication and social language use on the internet, such as postings on various social media platforms. Most social media platforms, such as Facebook or Reddit, do not limit the length of their postings to a meaningful degree. Platforms which do limit posting

length, such as Twitter, usually limit the maximum length, confining all of their content into the range of short texts, which is mathematically difficult to work with in quantitative corpus linguistics. Few online platforms require a minimum length for postings or even recommend postings to be of a specific length, whereas such requirements are commonplace within the publishing industry and for many of the genres included in typical published corpora, such as newspaper articles or academic writing. At the same time, online data has brought even the shortest texts to the center stage, making their societal and linguistic importance much more evident in comparison with the more traditional short genres. In other words, the free nature of internet writing has brought texts with a wide variety of lengths into the corpora of many linguists, and consequently made the problem of text length and, particularly, the problem of short texts more central than ever.

2.2 The importance of text length

It is clear that variation in text length, particularly the very shortest texts, cause mathematical issues in quantitative corpus-linguistic analyses. But do we actually have to care about text length? Can't we simply ignore the problematic cases when conducting quantitative linguistic studies? Or should we work towards finding more ways to make it possible to include texts of all lengths in our analyses?

Liimatta (2022a, 2022b) provides some insight into the role text length plays in linguistic variation by analyzing functional variation between comments of different lengths on the social media platform Reddit, with the very shortest comments also included in the analysis. In order to explain how these studies work around the problems caused by text length, it is worth describing the study design in some detail. To study functional variation, Liimatta (2022a, 2022b) builds on the ideas of multi-dimensional register analysis (MDA; e.g. Biber, 1988; Biber

& Conrad, 2009; Conrad & Biber, 2001) and more widely on the so-called text-linguistic framework of register analysis (cf. Biber et al., 2020). This framework is based on the idea that linguistic features are functional, and therefore tend to be used more in texts for whose situational context and communicative purpose they are better suited, and less when the situation or function of the text does not call for their use. Consequently, it is possible to measure variation in various linguistic features and use the observed variation patterns to suggest differences in the functions of texts. The features analyzed by Liimatta (2022a, 2022b) are based on the set of functional linguistic features used by e.g. Biber (1988) in studies of register variation.

In order to specifically focus on the functional variation taking place across text lengths, Liimatta (2022a, 2022b) makes use of so-called *lengthwise analysis*. This family of analysis methods aims to enable the analysis of linguistic variation between texts of different lengths, while at the same time circumventing many of the problems caused by variation in text length when using more typical corpus-linguistic methods. Lengthwise analysis is based on a simple insight: texts of different lengths may be difficult to compare with each other due to e.g. mathematical reasons, as explained above, but texts which are the exact same length can typically be compared trivially. For instance, it can be difficult to say just how comparable the rates of occurrence calculated from texts of different lengths are, but it is always possible to say how different two texts of the same length are in terms of e.g. the number of occurrences of some feature. A lengthwise analysis method, then, first performs a comparison between texts of the exact same length using some suitable method, and only after this *intra-length* comparison compares the results between texts of different lengths or *inter-length*.

Liimatta (2022a, 2022b) makes use of a simple but powerful lengthwise method for demonstrating the role of text length within Reddit. In the first-step intra-length comparison, the frequencies of various linguistic features were pooled and averaged by text length. In the

second step, the averaged frequencies were plotted across text lengths in order to show the variation in the feature frequencies across text lengths.

This kind of analysis requires a very large dataset, as there need to be enough texts of every length in the dataset for meaningful results. Consequently, social media is a good source of data for this method. Reddit in particular is arguably a very fruitful source of material for quantitative linguistic analyses overall, and especially for the analysis of the effects of variation in text length. First of all, Reddit enables access to large amounts of publicly available textual data. Some other social media platforms, such as Facebook, theoretically also have a lot of data available, but in practice a large portion of it is visible only to one's friends on the platform or those who have joined any specific discussion group. In other cases, the data may be public but difficult to access in large quantities in practice. Furthermore, since Reddit is divided into topic-based subforums called *subreddits*, the data is naturally subdivided into subcategories by topic and by register (see e.g. Liimatta, 2019). This allows studies to either focus on specific topic areas or draw comparisons between multiple ones. But most importantly when it comes to analyzing the effects of variation in text length, the comment length on Reddit is not limited for most practical purposes, allowing for comments of a wide range of lengths to exist, from extremely short to reasonably long. In a sense, one weakness of social media data for typical linguistic analyses becomes its strength in the analysis of variation across text lengths.

The analysis conducted by Liimatta (2022a) shows clearly that texts of different lengths play different roles on Reddit. The frequencies of various functional linguistic features vary across comment lengths: many of the features occur at different rates in Reddit comments of different lengths. This variation means that on Reddit as a whole, texts of different lengths have different functions. For instance, the results show that shorter Reddit comments include more features which are indicative of a more casual, interpersonal, and less edited style, whereas longer comments include more features which are considered more informational, edited, and

narrative. The differences in the comment functions are also shown to be the greatest between the very shortest comment lengths, with more gradual differences between longer comments.

Liimatta (2022b) performs a similar analysis but zooms in further to focus on a number of popular subreddits to find out whether all subreddits follow similar patterns, or if the same text length can have different functions in different subreddits. In the analysis, most subreddits analyzed are shown to follow similar patterns with each other. For example, the short comments in most subreddits contain more features which are more casual and involved, and longer comments contain more informational features. Similarly, comments of all lengths appear to be roughly equally narrative in all of the subreddits included in the analysis. However, a handful of the analyzed subreddits, which are more focused in terms of their topic in comparison with the very relative topics of most of the included subreddits, often differ greatly from both the general patterns and from each other. For instance, in the AskReddit subreddit, longer comments are much more narrative than the shorter ones, whereas for some other subreddits the opposite is the case. Figure 1 is an example of these results for past tense verb forms, which have been associated with narrative concerns (e.g. Biber, 2014). Figure 1 shows the pattern mentioned above: the frequency of past tense forms in most analyzed subreddits, shown in gray, stays fairly similar throughout the length range or even has a slightly decreasing pattern, pointing towards a relatively level spread of narrative concerns throughout the length range. However, the AskReddit subreddit in particular shows a dramatic increase in past tense verb forms as the comments get longer, implying that the longer comments in this subreddit tend to be on average increasingly narrative.

These results show that not only does text length play a role in linguistic variation, but that text length can be associated with functions differently within different register categories.

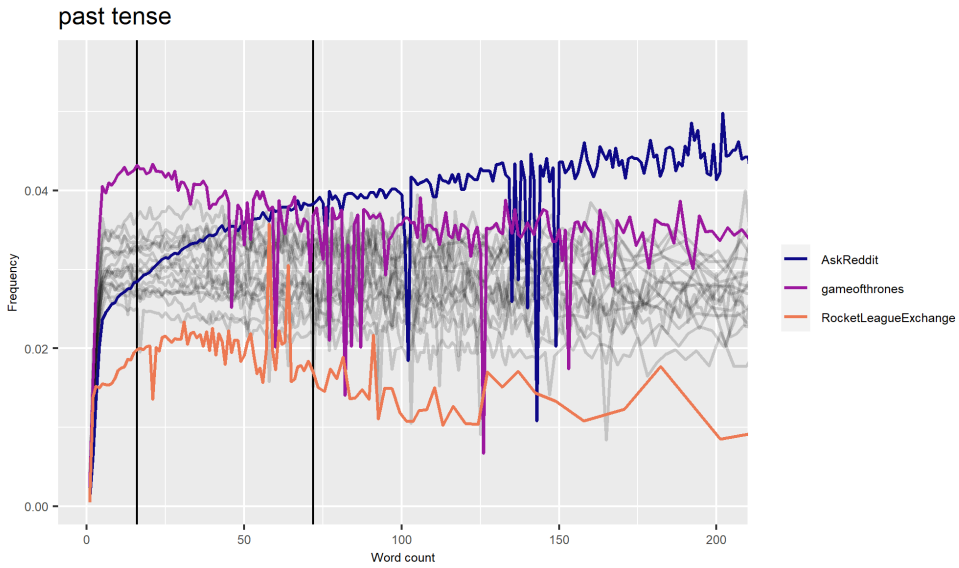

Figure 1. Frequency of past tense forms across comment lengths in 20 subreddits. From Liimatta (2022b: 279).

Taken together, the results of these two studies make it clear that texts of all lengths are of interest. If we ignore any of our dataset simply based on its length, because it is problematic in terms of quantitative analysis, we run the risk of excluding some of the variation within our data from our analysis. Consequently, the development of methods and approaches for the analysis of texts of all lengths is definitely a worthy endeavor, and potentially much more necessary than has been recognized before, particularly with the rise of sources of linguistic data including social media and other computer-mediated communication.

3. Solutions and workarounds

As the problem of text length and short texts has been recognized, a number of solutions and workarounds for it have also been devised. In this section, I will cover some of these approaches, and some related methods. Some of these approaches help solve or work around

the problems caused by variation in text length, some the problem of short texts, and some can help alleviate the effects of both. Additionally, I will describe a closely related problem, that of measures of lexical diversity. While the solutions to the problems with lexical diversity measures are not directly applicable to the problem of text length, they may still provide inspiration and starting points for new ways of approaching the problem of text length.

I have divided the solutions and workarounds to the problem of text length and short texts into two main categories. In the first group of approaches, the original set of texts is manipulated in some way, after which standard methods are applied. These could also be considered the more "traditional" approaches to the problem. Their advantage is that they are simpler to implement, but this means that their downsides are often greater. Conversely, the approaches in the second group make use of various statistical and/or computational methods to see the existing data in a new light. These approaches are more complicated to implement and often only work for specific kinds of analyses, but they are much more powerful in the situations for which they are well-suited. These two groups of course overlap in practice, and methods within and between the groups can even be used together.

3.1 Manipulation of the data

3.1.1 Exclusion

A commonly used workaround for the problem of short texts is to simply exclude all texts shorter than some threshold from the analysis. For instance, we might simply choose to remove all texts shorter than e.g. 400 words, 500 words, or 1,000 words from our dataset. If the aim is to be able to include as much of the data in our analysis as possible, this approach is at its most reasonable when there are only a small number of outliers under the chosen length limit, as the exclusion of a handful of outliers do not affect the overall results from a good-sized corpus very

much. It could even be argued that clear outliers do not even represent the varieties of interest in the corpus particularly well, and that therefore it would even be beneficial to exclude them.

However, the larger the proportion of the texts in the corpus which fall under the chosen length limit, the more problematic the exclusion method becomes. Particularly when typical texts from the shorter end of the length range start to be excluded from the analysis in addition to obvious outliers, it is clear that the dataset starts to lose some of the information potentially available within the data. Of course, it is not obvious where the line should be drawn when considering whether a text of a certain length should be considered an outlier, but from the point of view of the data, the best practice would be to have the length limit as low as possible. The optimal cutoff length when using the exclusion approach would be low enough that as many texts as possible are included in the analysis, but high enough that the desired analysis is still possible to conduct reliably. For a slightly more statistically-based approach than simply choosing some round number such as 400 or 500 as the limit, it is also possible to define the cutoff point as e.g. the 1% quantile of the length distribution, or whichever percentage gives a length limit which is workable with the chosen methods and the research questions being investigated, since this helps quantify the amount of data which has been left out.

However, there are datasets for which the exclusion approach is utterly unsuitable. For instance, most social media postings are very short, and therefore would need to be excluded from the analysis under any commonly used cutoff length which would allow analysis of the data using typical analysis methods. Consequently, different solutions and workarounds to the problems of short texts and variation in text length need to be used when dealing with such data.

3.1.2 Combining

In situations where discarding any data is undesirable, another workaround for the problem of

short texts is available. In many studies working with e.g. social media data and other genres which have a relatively large proportion of shorter texts, texts deemed too short to comfortably conduct the intended analysis on are combined together to create new "texts" which are sufficiently long for the analysis. For instance, we could decide to combine texts so that each of the combined texts is over some length limit, such as 500 or 1,000 words. As with the exclusion approach, the desirable length depends on the methods being used for the analysis and the research questions being investigated.

The main upside of the combining method, when compared to the exclusion method, is that no data is completely ignored: all text available for the analysis is included in the analysis. However, the downside is that by combining texts together, the texts lose their individual nature. For example, if one text is highly edited in style, and another one is highly casual, combining them together results in a loss of a lot of this information, and makes the combined text look somewhat average on both counts. In this way, the combining approach to the problem of short texts may very easily blur out some of the variation in the data. On the other hand, it is also possible that texts may end up combined in such a way that the resulting dataset overstates the importance of some feature which is actually quite rare overall. for instance if a feature is highly frequent in a small number of texts. The combining approach also easily results in a violation of the "independence assumption" inherent in various statistical procedures, including those commonly used by corpus linguists, such as Chi-square testing (Winter & Grice, 2021), and even precludes the use of some more advanced statistical methods such as the calculation of dispersion measures.

The issue of blurring out or overstating variation can in some situations be mitigated by the choice of the basis of combination. If the texts which are combined together are chosen in an essentially random manner, as is often the case, these potential obscuring effects cannot be reduced. However, in many cases, it is possible to use a more principled basis for the combining.

In the simplest case, the texts are combined based on some metadata in which we are interested in our analysis. For instance, if the analysis focuses on texts written by different sociolinguistic groups, any combining of texts needs to be done by the sociolinguistic groups in question. This, however, is done out of necessity, and it does not really help to reduce the blurring of variation taking place within the groups. For example, if we are comparing personal letters and official letters, we of course need to combine the shorter texts separately within the two categories, personal letters and official letters. But even in this case there may be variation within these two categories which gets either blurred out or overstated. In order to lessen the blurring and overstating effect, it might be useful to consider combining texts which are as similar as possible in their production circumstances, as far as reasonably possible. What texts exactly are considered "similar" is however a question which depends on the dataset and research questions. As a rule of thumb, however, the highest number of matching or similar metadata field values might be a good starting point.

Raising the level of analysis might be considered a special case of the principled metadata-based combining approach. For instance, instead of studying individual classified advertisements, we might consider the entire classified advertisements section a single text for the purposes of our analysis. Or instead of focusing on individual social media comments, we might choose to focus on full comment threads. The line between raising the level of analysis and the more general approach of principled metadata-based combining becomes blurred, however, in the case of e.g. combining Twitter tweets with their replies together to form individual texts.

3.1.3 Chunking

A different, slightly less-used approach to dealing with variation in text length is the opposite of combining shorter texts together: to cut longer texts into shorter pieces of (near) equal

length. For instance, Hiltunen & Tyrkkö (2019) make use of this approach when studying Wikipedia articles, which are extremely variable in length, by dividing the articles into 200-word pieces for their analysis.

When using this method, the fact that all texts included in the analysis are of (roughly) the same length facilitates their comparison using feature counts or rates of occurrence, since the confounding effects of variation in text length have been diminished. At the same time, all of the textual information is included in the analysis and not discarded, even if it has been cut into smaller pieces.

Texts can be split up in various ways. A straightforward approach is to simply split a text into chunks of a certain number of words. However, since sentences are a basic structural unit of language, placing chunk boundaries at sentence boundaries, making sure that every chunk includes enough words, is likely to be a more desirable solution in many cases. Another solution, which keeps the structural and discourse units of a text together even more, is to divide the text into its paragraphs, or multi-paragraph chunks.

In addition to the simple chunking options above, chunking can also make use of various computational methods to create chunks which are meaningful in terms of the discourse structure. For example, Biber et al. (2004) use a computational approach to divide texts into so-called Vocabulary-Based Discourse Units (VBDUs) in an analysis of the structure of various academic registers. VBDUs are a vocabulary-based approach to segmenting a text into discourse units. The methodology behind VBDU might be useful as a chunking approach in many different kinds of analyses into linguistic variation.

The chunking approach may or may not help with the problem of short texts. It would be difficult to meaningfully divide the longer texts into chunks of equivalent length if the shortest texts in the dataset are very short, such as on social media. On the other hand, if the shortest texts are still of reasonable length, dividing the longer texts into chunks of similar length might

actually make them more easily comparable.

3.2 Computational and statistical approaches

3.2.1 Lengthwise analysis

In order to make feature frequencies more comparable across text lengths, Liimatta (2020) proposes a family of methods called *lengthwise scaling*. Closely related to the lengthwise analysis described above, this family of methods is also based on the idea that it is trivial to compare texts which are the exact same length. In lengthwise scaling methods, feature counts in each text are first compared against texts of the exact same length ("intra-length comparison") using some suitable method of comparison. Based on the results of this comparison, each text receives a new, scaled value, which is a representation of how typical the text is in terms of the range of variation seen in all texts of the exact same length. These scaled values can then be compared between text lengths like normalized frequencies would be, such as by visual exploration of graphs or by using some further statistical or computational analysis.

While the idea behind lengthwise scaling can be applied in various ways, Liimatta (2020) demonstrates the method family with two specific implementations, *lengthwise rarity scaling* and *lengthwise quantile scaling*. In lengthwise rarity scaling, when computing the scaled value for a feature count, each feature count is compared against the full set of feature counts in texts of the same length. Each feature count is then replaced with the percentage of smaller feature counts in texts of the same length. In other words, the following question is asked for each text: "what percentage of all of the texts of the same length as this text has fewer instances of this feature?"

Of the two implementations, lengthwise rarity scaling is noted by Liimatta (2020) to be particularly useful for visual exploration of data. The advantages of this scaling method include

the fact that it particularly highlights smaller differences in rates of occurrence within the data, making it easier to pick up on subtler differences between groups of texts, and that it scales the observed variation into a constrained range between 0% and 100%, facilitating the graphing and interpretation of the results.

Figures 2 and 3 demonstrate the effect of lengthwise rarity scaling. Figure 2 is based on the typical method of calculating normalized frequency. It shows a kernel density estimation (KDE) of first-person singular pronouns in comments from three different subreddits sampled so that all three subreddits and different text lengths are represented more evenly in the figure in order to highlight the effects of the scaling method. The frequencies are clearly packed in the lower end of the frequency range, and it can be difficult to tell what is going on in this picture because of the heavy overlap. On the other hand, in Figure 3, the first-person singular pronoun counts have instead been scaled using the proposed lengthwise rarity scaling method. This scaling has created a much more clear differentiation between the use of first-person singular pronouns in the three subreddits.

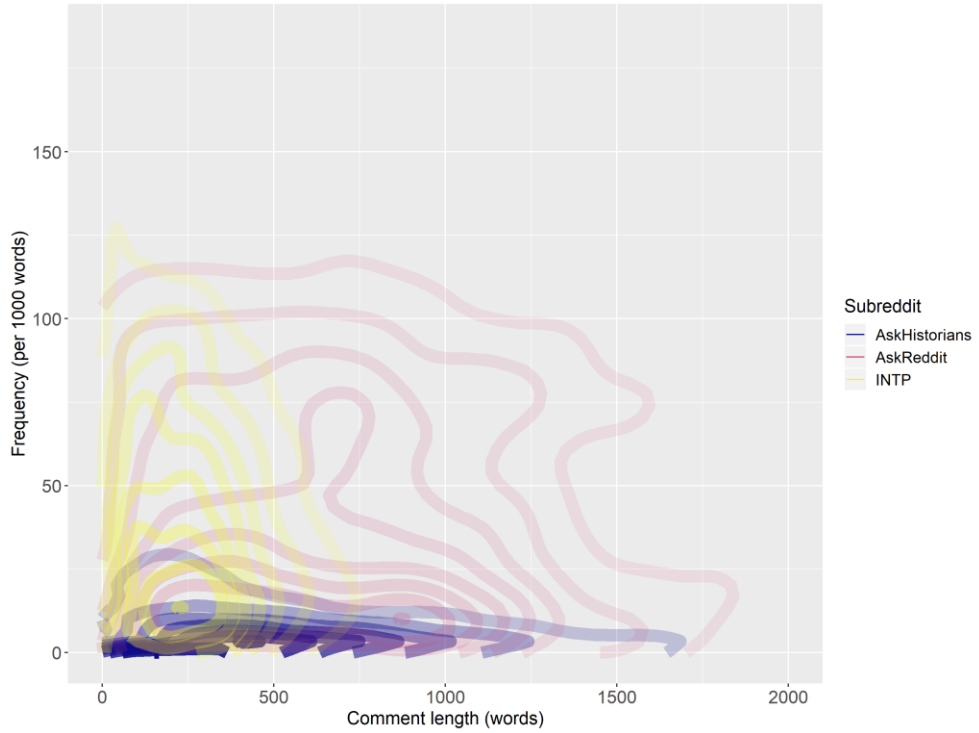

Figure 2. Kernel density estimation of a sample of first-person singular pronoun frequencies

across different comment lengths with contour lines for 8 bins. From Liimatta (2020: 121).

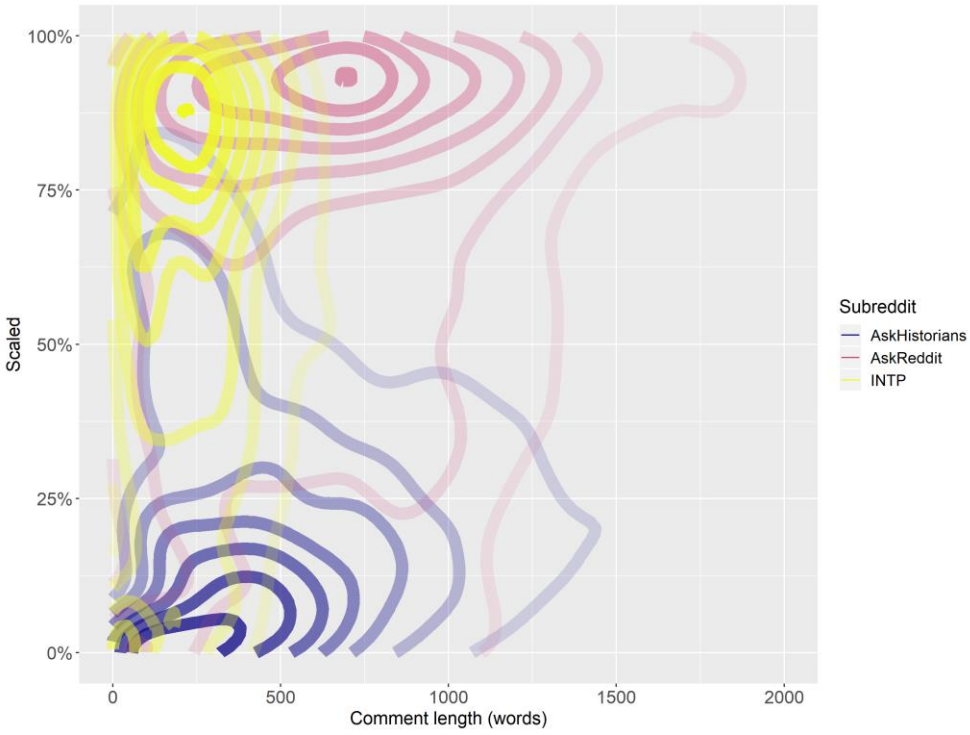

Figure 3. Kernel density estimation of a sample of lengthwise rarity scaled first-person singular pronoun counts across different comment lengths, divided into 8 bins. From Liimatta (2020: 123).

The second implementation of the lengthwise scaling method family proposed by Liimatta (2020), lengthwise quantile scaling, builds on commonly-used statistical properties of the dataset. In particular, the scaled scores within each text length are scaled based on the median feature count and specific quantiles of the feature count distribution. In the proposed implementation, all feature counts are scaled so that the median feature count within each text length becomes 0, and the values above and below the median are scaled separately from each other so that the .05 quantile becomes -1 and the .95 quantile becomes 1.

Since lengthwise quantile scaling is based on the median and certain quantiles of the data, the -1 and 1 lines are particularly useful for the interpretation of the data in terms of

recognizing texts with uncharacteristically high or low feature counts for any given text length. On the other hand, in contrast with lengthwise rarity scaling, lengthwise quantile scaling does not confine the scaled values into any particular range. It also does not highlight smaller differences as well as lengthwise rarity scaling. However, thanks to its basis on common statistical measures, lengthwise quantile scaling may be the better of the two methods to use as a preprocessing step for further statistical or computational analysis.

The main downside to both lengthwise rarity scaling and lengthwise quantile scaling is that they require a very large dataset, so that there are enough texts of every individual length to make it possible to compare texts of the same length. Such datasets are most readily based on social media and other online sources. However, if the dataset mostly contains longer texts, even a slightly smaller dataset will do if texts of adjacent lengths are binned together. If the dataset is even smaller still, and/or includes shorter texts as well, the two methods may not work too well. But these two methods are only two potential implementations of the lengthwise scaling method family. In situations where the dataset is relatively small and includes a large number of shorter texts, other kinds of implementations of the lengthwise scaling method family may work better. For instance, Liimatta (2020) suggests the use of resampling methods, which could be used even with a smaller corpus, and possibly in conjunction with e.g. binning, to estimate the population parameters for different text lengths, which can then be used as the basis for the comparison.

However, even this method is unlikely to work with the smallest corpora and the shortest texts, which do not have enough text for each text range to estimate the population parameters with any reliability.

3.2.2 Multiple Correspondence Analysis

There also exist methods for specific purposes which can be used with shorter texts. For

instance, factor analysis methods, such as those used in the multi-dimensional method of register analysis, rely on feature frequencies, and as such the methodology is difficult to apply to genres which include a large proportion of short texts. In order to get around this issue in their multi-dimensional analyses of Twitter tweets, Clarke & Grieve (2017, 2019) make use of a method called *multiple correspondence analysis* (MCA). MCA is a dimensionality reduction method which can be used to extract dimensions of variation from a set of variables. However, unlike methods such as factor analysis or principal component analysis, which rely on continuous variables such as feature frequencies, MCA extracts its dimensions from categorical variables, such as the presence or absence of a feature in a text. Consequently, MCA can be used to analyze the dimensions of variation within genres with extremely short texts, such as tweets.

However, while MCA works well with genres with only short texts, it cannot be used with datasets which include longer texts. This is because the longer a text becomes, the more likely it is to include any given feature. As the texts get longer, more and more of the features of interest start appearing in every text. Due to this, the co-occurrence patterns end up saturated when analyzing longer texts, rendering the method unusable with such texts.

3.2.3 Resampling methods

Resampling methods are powerful statistical methods which "make the best use of the available data" (Säily, 2014: 47) and create confidence intervals and estimate the population parameters for e.g. the rate of occurrence of a linguistic feature or item. Resampling methods have been developed with the idea that the dataset is only a sample, an imperfect representation of the full population it is supposed to represent. When using resampling methods, the dataset is first divided into samples, e.g. texts, groups of texts, or parts of texts, depending on the research questions and statistical assumptions being made. Then, this set of samples is repeatedly sampled randomly in order to create new artificial datasets. The exact details of how the

sampling is conducted depend on the specific resampling method chosen, such as permutation or bootstrapping. The results of these different methods also have different interpretations. The new set of artificial datasets is then analyzed in order to estimate the value and its confidence intervals.

Resampling methods have been used in various studies of linguistic variation. They can be used simply to estimate the rate of occurrence together with its confidence intervals, or to enable analysis in situations where using the standard method of normalization is difficult (e.g. Gries, 2006, 2022; Lijffijt et al., 2016; Säily, 2014). While studies making use of resampling methods often do not explicitly deal with the problem of text length specifically, the core idea of the methodology is very applicable to this problem as well.

3.3 A related problem: lexical diversity

While the effects of text length have generally speaking not been studied very much in corpus-linguistic research, there is a group of measures, whose relationship with text length has received some more attention: the *type-token ratio* and other measures of lexical diversity (or "lexical richness"). While the type-token ratio differs as a measure from the typical calculated normalized frequencies, its relationship to text length still bears discussing in this context.

Like its name suggests, the type-token ratio is the ratio of the number of different words in a text (types) to the number of all words in the text (tokens). This ratio is notoriously sensitive to variation in the length of the text it is calculated for. Due to this sensitivity, for the results to be comparable, the ratio should be calculated for texts of almost the exact same length. However, since all texts in a normal-sized corpus are rarely close enough to each other in length, as a typical workaround, the ratio is calculated for a set number of words (such as 400 words) taken from the beginning of each text. While this workaround has been used for a long time to

good effect, it is also not optimal, since in many cases it excludes a large majority of the text from the calculation. The solution is a lot less optimal still for datasets with a lot of variation in text length, since the 400 word sample covers a different fraction of each text, which means that every text is represented differently by the sampling.

Due to these problems, and the fact that being able to measure lexical diversity in a meaningful way would be very desirable for many linguistic questions, the question of whether a method which is less sensitive to text length could be devised has received a decent amount of attention from corpus linguists and others. Hess et al. (1986) and Hess et al. (1989) test various mathematical transformations of the basic type-token ratio, and conclude based on their results that no simple transformation can make the type-token ratio comparable across text lengths. Because of the inherent problems with the type-token ratio, various alternative measures of lexical diversity have been created. These include, for example, the Moving-Average Type-Token Ratio (MATTR) (Covington & McFall, 2010) and the Moving Window Type-Token Ratio Distribution (MWTTRD) (Kubát & Milička, 2013). Some other methods of calculating a lexical diversity score can be found in e.g. Koizumi & In'nami (2012), who compare the performance of six different measures of lexical diversity in shorter text samples between 50 and 200 words, and in Shi & Lei (2020), who more recently compare two entropy-based measures of lexical diversity.

The problem of lexical diversity measures is closely related to the problem of text length and short texts in focus in the present study. While the efforts to develop a measure of lexical diversity which is less affected by text length do not directly target the problem of text length and short texts, the implication of these efforts is clear: methods which lessen the confounding effects of variation in text length can be developed. Maybe some method created for the purpose of measuring lexical diversity could even be adapted to help with the problem of text length in feature frequencies.

4. Conclusion

The present chapter has discussed two related problems, the more general problem of variation in text length and the more specific problem of short texts. While these problems have not received as much attention than they could have from quantitative corpus linguists (as evidenced by e.g. the body of research on measures of lexical diversity), the difficulties caused by the confounding effects of text length are only going to become more central to many studies, as more and more research is being done on social media and web data.

A number of solutions and workarounds to remedy the problems have been devised, all with their own advantages and disadvantages. These solutions can be used to good effect in many kinds of linguistic investigations. However, there still is no one-size-fits-all solution to the problems caused by text length and short texts in quantitative text-analytic corpus-linguistic studies. Some potential avenues for improvements and new method development have been proposed in the present chapter.

Since resampling methods are very powerful for estimating the distribution based on smaller datasets, they appear as a potentially useful avenue for the development of new methods for the analysis of texts across text lengths. At the same time, larger datasets contain more information about the variation inside them, so various approaches making use of the large size of the data, such as those developed by Liimatta (2020), may also be useful in getting around many of the problems caused by text length in at least some studies. However, such approaches can naturally only be used with a limited number of datasets which are large enough.

Even if a perfect all-encompassing solution does not exist yet, or is not possible at all, the

solutions mentioned in this chapter can still be used to study many linguistic questions, given that one is aware of the potential implications of their use. There certainly exist many other approaches not mentioned here, particularly various more advanced statistical and computational methods, which are less affected by variation in text length. Nevertheless, there is still a lot of room left for the development of new ways to analyze datasets with a wide range of text lengths, and particularly datasets which contain extremely short texts, which are more common today than ever.

References

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in contrast, 14*(1), 7-34. https://doi.org/10.1075/lic.14.1.02bib

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press. https://doi.org/10.1017/CBO9780511814358

Biber, D., Csomay, E., Jones, J. K., & Keck, C. (2004). A corpus linguistic investigation of vocabulary-based discourse units in university registers. In U. Connor & T. A. Upton (Eds.), *Applied Corpus Linguistics: A Multidimensional Perspective* (pp. 53-72). Rodopi.

Biber, D., Egbert, J., & Keller, D. (2020). Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory, 16*(3), 581-616. https://doi.org/10.1515/cllt-2018-0086

Clarke, I., & Grieve, J. (2017). Dimensions of abusive language on Twitter. In Z. Waseem, W. Hui Kyong, D. Hovy, & J. Tetreault (Eds.), *Proceedings of the First Workshop on Abusive Language Online* (pp. 1-10). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3001

Clarke, I., & Grieve, J. (2019). Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PLoS ONE, 14*(9). https://doi.org/10.1371/journal.pone.0222062

Conrad, S., & Biber, D. (Eds.). (2001). *Variation in English: Multi-dimensional studies*. Pearson Education. https://doi.org/10.4324/9781315840888

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94-100. https://doi.org/10.1080/09296171003643098

Gries, S. T. (2006). Exploring variability within and between corpora: Some methodological considerations. *Corpora, 1*(2), 109-151.

Gries, S. T. (2022). Toward more careful corpus statistics: uncertainty estimates for frequencies, dispersions, association measures, and more. *Research Methods in Applied Linguistics, 1*(1). https://doi.org/10.1016/j.rmal.2021.100002

Hess, C. W., Haug, H. T., & Landry, R. G. (1989). The reliability of type-token ratios for the oral language of school age children. *Journal of Speech and Hearing Research, 32*, 536-540. https://doi.org/10.1044/jshr.3203.536

Hess, C. W., Sefton, K. M., & Landry, R. G. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech and Hearing Research, 29*, 129-134. https://doi.org/10.1044/jshr.2901.129

Hiltunen, T., & Tyrkkö, J. (2019). Academic vocabulary in Wikipedia articles: Frequency and dispersion in uneven datasets. In C. Suhr, T. Nevalainen, & I. Taavitsainen (Eds.), *From Data to Evidence in English Language Research* (pp. 282-306). Brill.

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System, 40*(4), 554-564. https://doi.org/10.1016/j.system.2012.10.012

Kubát, M., & Milička, J. (2013). Vocabulary richness measure in genres. *Journal of Quantitative Linguistics, 20*(4), 339-349. https://doi.org/10.1080/09296174.2013.830552

Liimatta, A. (2019). Exploring register variation on Reddit: A multi-dimensional study of language use on a social media website. *Register Studies, 1*(2), 269-295. https://doi.org/10.1075/rs.18005.lii

Liimatta, A. (2020). Using lengthwise scaling to compare feature frequencies across text lengths on Reddit. In S. Rüdiger & D. Dayter (Eds.), *Corpus approaches to social media* (pp. 111-130). John Benjamins. https://doi.org/10.1075/scl.98.05lii

Liimatta, A. (2022a). Register variation across text lengths: Evidence from social media. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.20177.lii

Liimatta, A. (2022b). Do registers have different functions for text length? A case study of *Reddit*. *Register Studies, 4*(2), 263-287. https://doi.org/10.1075/rs.22007.lii

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., & Mannila, H. (2016). Significance testing of word frequencies in corpora. *Digital Scholarship in the humanities, 31*(2). https://doi.org/10.1093/llc/fqu06

Shi, Y., & Lei, L. (2020). Lexical richness and text length: An entropy-based perspective. *Journal of Quantitative Linguistics*, 1-18. https://doi.org/10.1080/09296174.2020.1766346

Säily, T. (2014). *Sociolinguistic variation in English derivational productivity: Studies and methods in diachronic corpus linguistics*. Société Néophilologique de Helsinki.

Winter, B., & Grice, M. (2021). Independence and generalizability in linguistics. *Linguistics, 59*(5), 1251-1277. https://doi.org/10.1515/ling-2019-0049