DRAFT

A statistical analysis of word-lengths in the Pali canon

Venerable Vimala Bhikkhunī

February 7, 2020

Contents

1 Summary

The Pali language, as every other language, has evolved over time. One of the changes that the language has undergone through the centuries is that words have become increasingly complex and words have merged together to form longer words. We would therefore expect that a statistical analysis of average word-length could be an indication of relative "age" of the texts in question.

In this research I analyse the average word-length of each text and each book in the Pali canon to find trends in the canon and test the above hypothesis; to see if average word-length can be used as an indicator to show the development of the texts over time.

I found that not only is there a trend in average word-length within the canon, but that there are some surprising anomalies that I investigated further.

Average word-length in itself remains only an indication, among other indications like the existence and quality of parallels, but nevertheless can prove a valueble resource in determining the development of the texts in the Pali canon and ultimately what the Buddha taught.

2 Introduction

3 Methodology

3.1 Sourcing texts

For this research, I used the Pali canon from the Mahāsaṅgīti Tipiṭaka Buddhavasse 2500 from the Vipassana Research Institute (`https://tipitaka.org/`).

These texts, with the exception of the commentarial texts, are used by SuttaCentral.net and they have partly segmented these files and coded them in json. These json files do not have any of the coding that is necessary to display texts online but just the pure texts, which makes it ideal to work with. This format is the basis I used for my analysis.

Where these files did not exist, I have taken them from the SuttaCentral HTML and coded them into the same segmented json format. The same I have done for the commentarial texts directly from the VRI website in XML.

The github repository for these json files is here: (`https://github.com/BuddhaNexus/segmented-pali`)

3.1.1 Variant readings

The Mahāsaṅgīti Tipiṭaka as used by the VRI as well as SuttaCentral also list various variant readings. These were removed for the purpose of this analysis. After analysing part of the variant readings I concluded that these would not make a significant shift in the calculations. In some cases variant words were a bit longer, in other cases shorter.

3.1.2 Abbreviations

Many texts contain abbreviations in the form of "⋯pe ⋯" or similar. These abbreviations were removed prior to calculations because they would give a much lower average word-length for texts with many abbreviations. Of course this will go from the premise that the text being substituted has an overall average that is roughly the same as the still remaining text. I feel this is reasonable assumption and above all, it would be undo-able to replace the abbreviations with the text they are to replace.

3.2 Other materials used

In some instances I had to look into the parallels distribution of files within a collection. For this I used the parallels json file that is used by SuttaCentral. Although this material is not complete i.e. these are human sourced parallels and not all parallels will be known, they give an indication. Moreover, these parallels also list known parallels with other Buddhist canons in other languages.

One major drawback in using this parallels-list for statistical analysis is that it is not consistent in how it counts parallels. For instance text A can be parallel to text B and

this can list as one parallel. But it is also possible that a more detailed listing is applied in that for instance text A, verse 1 is parallel to text B, verse 2, etc. This multiplies the number of parallels considerably.

In analysis some collections, I also made use of a sankey map and other graphs as used by BuddhaNexus.net. Although this is very detailed and complete, it only lists matches within the same canon. Matches within the same canon have their value, but in matching between various languages we can determine if texts were already in existence before the split of the schools and therefore their relative age (see [**?**]). The matching between languages is in the pipeline but to date not yet available.

Note also that BuddhaNexus graphs show the total lengths of matches so that a long match between texts counts for more than just one line, as opposed to the parallels listed in SuttaCentral, where every match just counts as 1. Therefore, it also takes the quality of the match into account.

3.3 Method of analysis

The segmented files were further imported into an ArangoDB database and queries made from there, using the Python programming language for retrieving meaningful data for barcharts.

All punctuation and other typography markers and numbers were removed so only the alphanumerical characters were used for calculating the average word-length.

4 Average Word-length analysis

4.1 First results

After calculating and plotting the average word-length of all collections in the Pali canon, the following chart ensued:

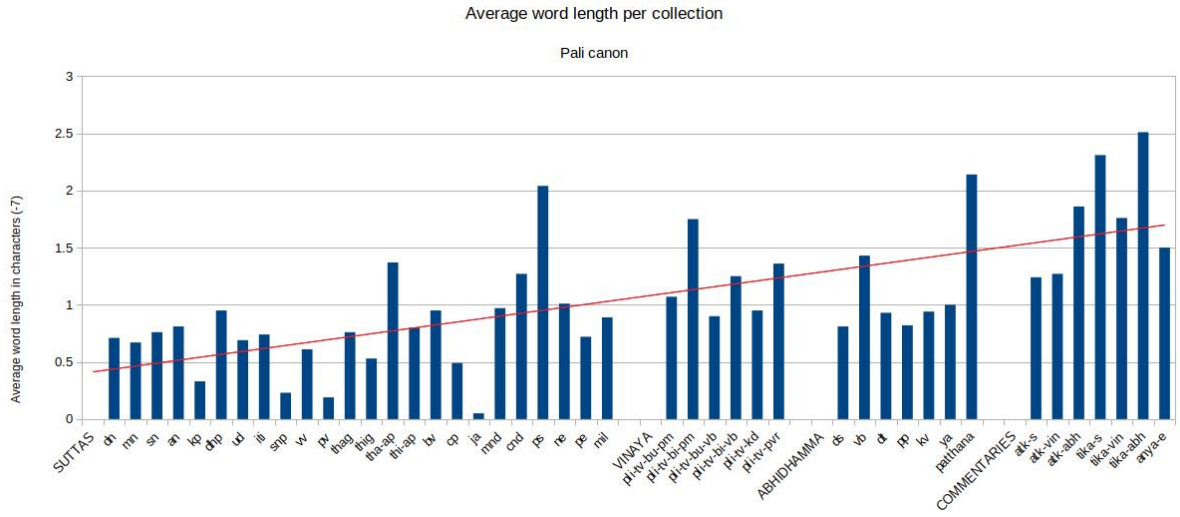

Figure 1: The average word-length in characters above 7 per book in the Pali canon. The red line shows the linear trendline (regression through equation y=a*x+b)

Some interesting things can be seen from this chart. First of all, it looks like those books that are generally accepted as "Early Buddhist" (see [**?**]) indeed have a lower average word-length while the later commentarial texts have a much higher average word-length.

The Vinaya and Abhidhamma texts as well as the later Sutta texts seem to all be in between with roughly the same average word-length, indicating that these texts might have developed at the same time.

However, there are also some books that do not seemingly match this broad pattern.

The first thing that struck me is the relative high value of the Dhammapada, which is generally seen as one of the earliest Buddhist texts. Then there are other discrepancies like the Jātaka and the Bhikkhuni Patimokkha. So I will explore these in more detail below.

4.2 Dhammapada

Analysing the Dhammapada (`https://suttacentral.net/dhp/pli/ms`), I noticed that especially the headings have a much larger word-length than the verses.

If we keep in mind that all early texts were only transmitted orally and organized and written down at a much later date, it is quite possible that headings were inserted into the texts later.

So calculating the same chart with the headers taken out gives the following interesting picture:

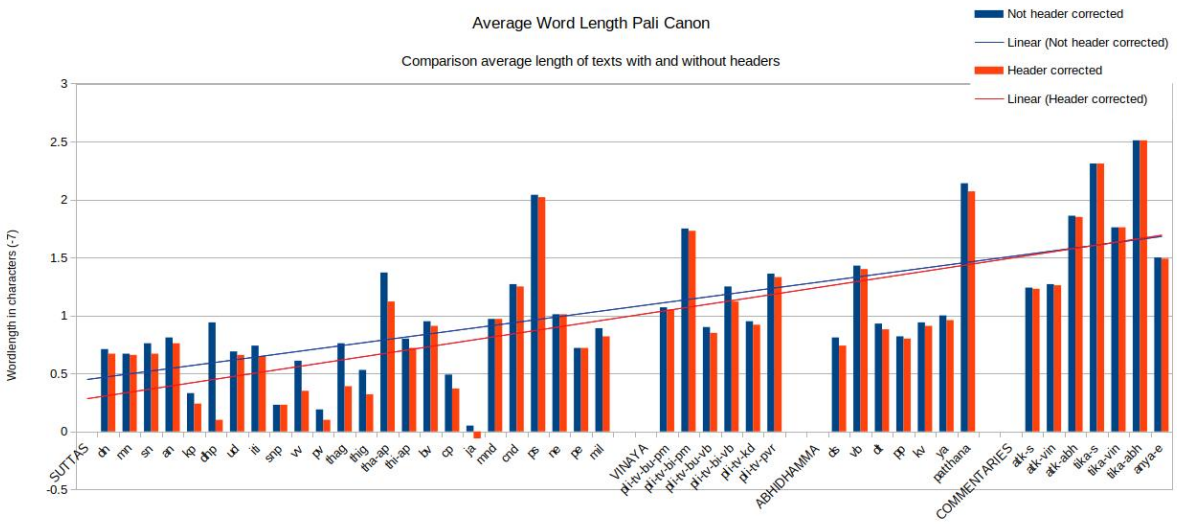

Figure 2: The average word-length in characters above 7 per book in the Pali canon comparing calculations with and without headers. The blue and red lines shows the linear trendlines resp. (regression through equation y=a*x+b)

Removing headers from the calculation and just analysing the prose or verse text suddenly gives a very different picture for expecially the Dhammapada, which has now jumped from a relatively high value to a much lower value which is much more in line with our accepted understanding of the Dhammapada as a very early text.

Another interesting thing we see in this chart is that removing the headers has a much more dramatic effect on the Early Buddhist texts and the commentarial texts are hardly affected. So I think it is safe to assume that headers are indeed inserted into the texts later.

4.3 Anguttara Nikaya

For comparison, the next chart shows the same data but with the 0-line at the level of the highest of the Early Buddhist books, which is the Anguttara Nikaya. The fact that the Anguttara Nikaya is the book with the highest average word-length is consistent with our understanding that of all the early Buddhist Nikayas, this book has had the most influence by later insertions.

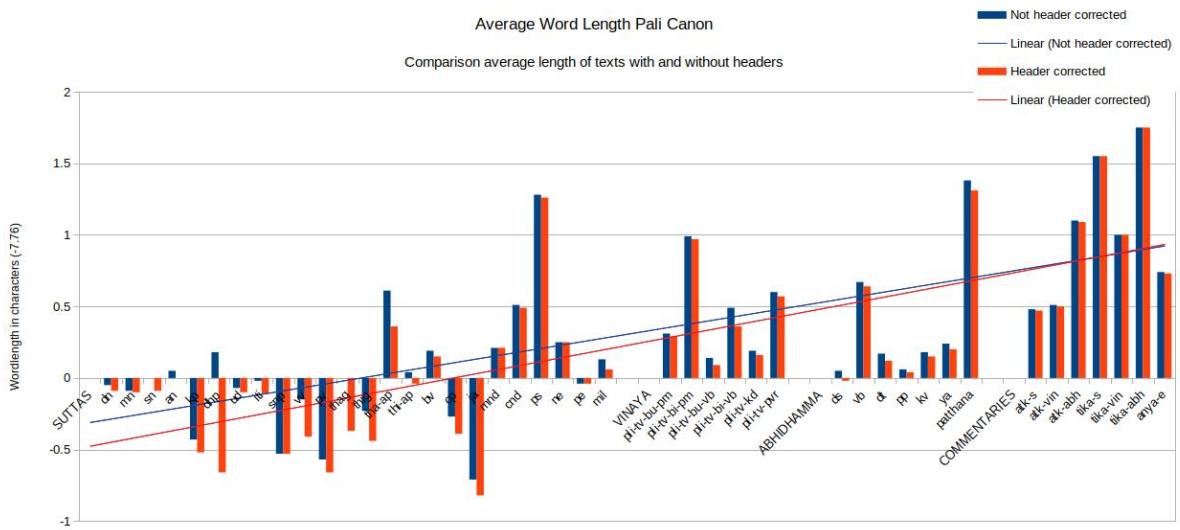

Figure 3: The average word-length in characters above 7.76 per book in the Pali canon comparing calculations with and without headers. The blue and red lines shows the linear trendlines resp. (regression through equation y=a*x+b)

4.4 Verse vs. prose

When we look at the early Buddhist books, from the Digha Nikaya through to the Therigatha, we notice another interesting phenomenon: books with predominantly verses have a much lower average word-length than books with prose.

There are various ways in which this can be explained. First of all, it might be that verse, by its very nature, needs shorter words. But we will see in the analysis of the Jātaka that this is not always the case.

Another possibility is that due to the history of oral tradition, it was much easier to remember and recite verses than prose and we do indeed see that verse summaries of prose teachings were one method to ensure oral transmission.

4.5 Bhikkhuni Patimokkha

The average word-length of the Bhikkhuni Patimokkha seems very high in comparison to the other books of the Vinaya. This is not surprising if we remember that the Bhikkhuni Patimokkha had been lost and was actually reconstructed from a commentarial text. No doubt due to it's history, the text would also have acquired some of it's later use of the Pali language, most notably with concatted and therefore longer words. This does not say anything about the actual age of the original Bhikkhuni Patimokkha, only of the current version we have.

4.6 Jātaka

One of the most striking anomaly is the Jātaka collection. In fact, it is the collection with by far the lowest average word-length.

The Jātaka verses are verses are generally seen as pre-Buddhist tales. They came into the Buddhist canon, not only because the Buddha himself sometimes used these tales in his teachings, but more because the old tales would get a story woven around them in which these were said the stories of the Buddha's previous lives. However, these stories are not in the Mahāsaṅgīti Tipiṭaka version of the Jātaka, where only the verses are used. So it is maybe not so surprising that these stem from an earlier use of the Pali language and therefore have a shorter overall word-length.

The below figure shows the distribution of the Jātakas.

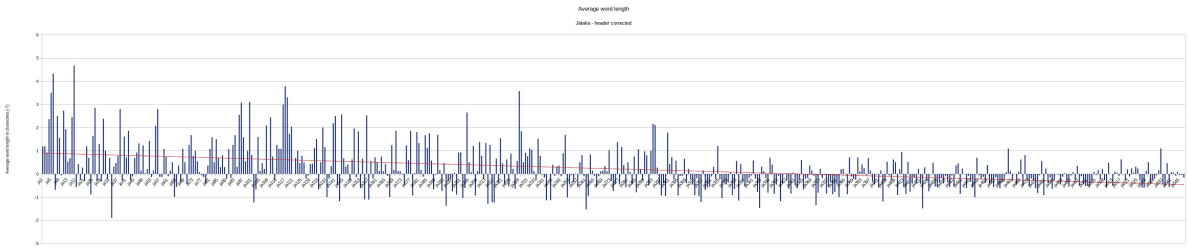

Figure 4: The average word-length in characters above 7 per book in the Pali canon comparing calculations with and without headers. The red line shows the linear trendline (regression through equation y=a*x+b)

This figure shows an interesting trend: the shorter Jātakas with a lower number seem to have a much higher word-length on average than the ones further in the collection.

The next figure shows the number of parallels listed on SuttaCentral for the texts in this collection on a logarithmic scale.

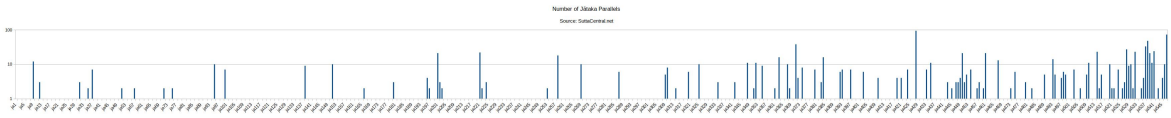

Figure 5: The number of parallels as listed on SuttaCentral on a logarithmic scale

Although there are not very many known parallels for many of these texts, it is clear from this that the most parallels appear in the higher end of the Jātaka collection.

From both of these charts it would seem that the Jātaka with a higher number are likely to be earlier than the first part of the collection. Of course we cannot make any assumptions about the lateness or earlyness of individual suttas based on this method, especially if many of these only contain just one verse, but it is an indication of a general trend.

Another interesting collection as is seen in figure 3 is the Cariyāpiṭaka. A small col-

lection of only 35 verse-texts, it mainly comprises of retellings of the Jātaka stories. Analysing this in a chart we get the following:

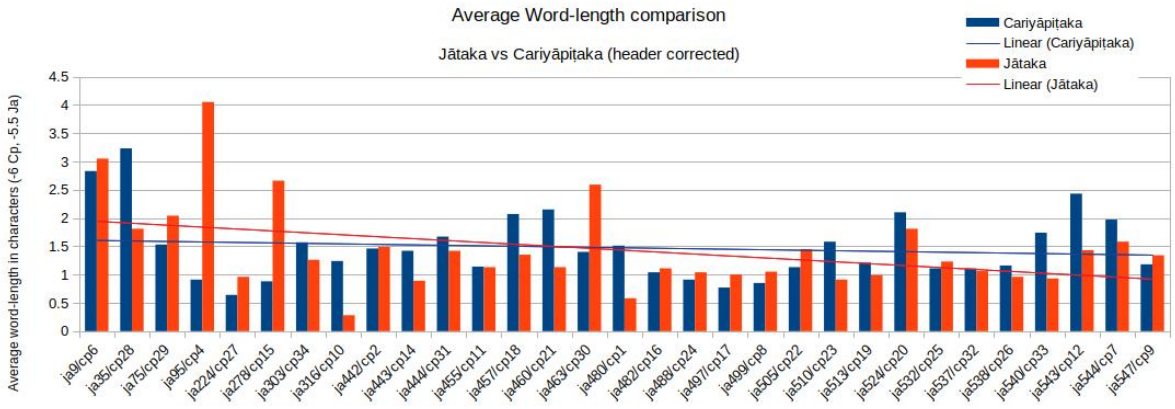

Figure 6: Comparison of Jātaka with Cariyāpiṭaka (Note that for the purpose of this comparison the value of length for Cariyāpiṭaka is .5 characters higher than for Jātaka)

There seems to be a rough, but not entirely convincing, similar trend between the two collections. More notably we see that most of the Cariyāpiṭaka are retellings of the Jātaka with higher numbers. This could be another indication that the latter part of the collection is earlier than the first part.