Toward a Computational Analysis of the Pali Canon

Dan Zigmond

Abstract

This paper describes the results of applying computational text mining to the *Tipiṭaka,* or Pali Canon, the canonical scripture of Therevāda Buddhism. Individual volumes of the *Tipiṭaka* are divided into "clusters" using purely computation tools, and in many cases these clusters appear to match the rough scholarly consensus around the relative age of the volumes. Texts are also summarized into "word clouds" based on relative word frequency, and these also seem to reflect the underlying themes of the texts. While these initial results are essentially confirmational rather than novel, they suggest these approaches will be valuable additions to the Pali scholar's toolbox.

Computational text mining

Text mining can be defined as the process of extracting new information from textual sources using computational means. Its practice goes back to the invention of digital computers in the mid-twentieth century, and the explosion of availability of both texts in electronic form and the computing capacity necessary to analyze them have greatly accelerated progress in recent years. Although much initial work was focused on commercial and government applications, the past decade has seen increasing adoption of computational techniques in the humanities (Jockers 2013).

JOCBS. 2021(20): 107–135.

Thus far advances in text mining have typically been applied to English and other modern (and primarily Western) languages, but this is starting to change. Since 2007 researchers have convened regular international meetings on Sanskrit Computational Linguistics (see, for example, Kulkarni and Dangarikar 2013). Along similar lines, the Classical Language Toolkit aims to make text mining applicable to many ancient languages (Johnson et al. 2014). To the best of our knowledge, however, there have been very few attempts to apply these techniques systematically to the Pali Canon (e.g., Elwert et al. 2015).

In recent years a robust set of generalized tools have emerged to support computational analysis, making application of these techniques to novel corpora and languages more feasible. The work in this paper was carried out using the *R* statistical programming language (R Core Team 2013), and the *tidy* (Wickham 2019), *tidytext* (Silge and Robinson 2016), *factoextra* (Kassambara and Mundt 2020), and *wordcloud* (Fellows 2018) packages.[1]

Particular challenges of the Pali Canon

The *Tipiṭaka,* or Pali Canon, is the canonical scripture of Therevāda Buddhism. Purported to be the oral teachings of the historical Buddha (Sujato & Brahmali 2014), it is believed to have been first recorded in written form in what is now Sri Lanka around the first century BCE. Although versions of these texts are preserved in other languages, the oldest and most complete edition of the *Tipiṭaka* is recorded in Pali, a Middle Indo-Aryan dialect whose name derives from the compound *pāli-bhāsa*, "the language of the texts" (Geiger 2005, xxiii). In other words, the Pali language and the *Tipiṭaka* are intimately linked: Pali is literally the language of the *Tipiṭaka*. As Pali scholar Richard Gombrich put it succinctly: "For many Buddhists, Pali occupies the kind of place that Arabic occupies for Muslims, Hebrew for Jews, [and] Greek and/or Latin for various kinds of Christians" (Gombrich 2018).

Pali appears to have been originally a spoken language and has no fixed written form. It has been traditionally written using the script of the various Asian countries where Therevāda Buddhism has proliferated: in Khom and Tham scripts in Thailand, Burmese in Burma, Sinhalese in Sri Lanka, etc. The growth of the Vipassana meditation movement founded by S.N. Goenka in India has led to a resurgence of Pali texts printed in Devanagari script. Beginning in

[1] The novel tools and electronic texts used in this paper are freely available through the *tipitaka* package (Zigmond 2020), and the source code for this package can be found at https://github.com/dangerzig/tipitaka.

the late 19th century, the Pali Text Society (PTS) pioneered the publication of Pali texts in Roman script for Western scholars (and Western Buddhists) using a system of diacritics similar to that typical for transliterated Sanskrit. This Roman rendering is the written form of Pali used in this analysis.

For example, here is the first verse of the *Dhammapada*, perhaps the most famous of the Pali scriptures and the first to be translated into a European language, first in Roman-scripted Pali, then in two modern translations:

> *manopubbaṅgamā dhammā, manoseṭṭhā manomayā,*
> *manasā ce paduṭṭhena bhāsati vā karoti vā,*
> *tato naṃ dukkham anveti, cakkaṃ va vahato padaṃ.* (Dhp 1)

> Preceded by perception is mental states,
> For them is perception supreme,
> From perception that have sprung,
> If, with perception polluted, one speaks or acts,
> Then suffering follows,
> As a wheel the draught ox's foot. (Carter and Palihawadana 1987)

> All experience is preceded by mind,
> Led by mind,
> Made by mind,
> Speak or act with a corrupted mind,
> And suffering follows
> As the wagon wheel follows the hoof of the ox. (Fronsdal 2005)

The title itself of the Pali Canon, *Tipiṭaka,* can be translated as "consisting of three baskets" and the Canon is composed of three distinct sets of scriptures:

- *Vinaya Piṭaka,* Basket of Discipline, describing the rules for the monastic order.
- *Sutta Piṭaka*, Basket of Discourses, primarily recounting the direct teachings of the Buddha (such as the verse quoted above).
- *Abhidhamma Piṭaka,* Basket of Special Teachings,[2] summarizing and systematizing the Buddha's doctrines.

[2] An alternative understanding of *Abhidhamma Piṭaka* would be the basket "about the teachings."

Each of these is composed of several books, which in turn are often divided into chapters and verses. The *Sutta Piṭaka* is the most widely studied and so its divisions have particular significance. It contains four major collections of *suttas* or discourses, plus a fifth collection of a wide variety of generally shorter material.

Table 1 shows these major divisions and the approximate length (in words) of each. The total size of the *Tipiṭaka* is just under 2.7 million words, with the *Suttas* alone totaling near 1.5 million. By way of comparison, the King James Version (KJV) of the Christian Bible contains approximately 855,317 words (Project Gutenberg 2020). Thus, in (very) rough terms, the *Tipiṭaka* (in Pali) is a bit more than three times the length of the KJV (in English), while the Buddha's discourses alone (i.e., the *Sutta Piṭaka*) are a bit less than twice the length of the KJV.

Components of the *Tipiṭaka*	**Approximate length (in words)**
Vinaya Piṭaka	**414,887**
Sutta Piṭaka	**1,475,446**
Dīgha Nikāya	142,313
Majjhima Nikāya	244,973
Saṃyutta Nikāya	264,973
Aṅguttara Nikāya	300,010
Khuddaka Nikāya	523,177
Abhidhamma Piṭaka	**801,650**
***Tipiṭaka* TOTAL**	**2,691,983**

*Table 1: Divisions of the **Tipiṭaka** and length in words*

There are a number of challenges to working with this material using computational tools. First there are several extant versions of the Canon. This analysis was based on the digital edition of the Chaṭṭha Saṅgāyana Tipiṭaka version 4.0 published by Goenka's Vipassana Research Institute (hereafter CST4; Vipassana Research Institute 2020). This edition originated at the so-called 'Sixth Buddhist Council', held in Burma from 1954 to 1956. Originally published after the Council meetings in Burmese script, the Vipassana Research Institute in India began printing this edition in Devanagari and eventually Roman (and several other) scripts in 1990 and later published the results electronically as well.

This CST4 edition differs somewhat from the more widely used Roman edition published by the PTS in the UK, although no exhaustive catalog of the inconsistencies appears to exist. While the PTS edition is available electronically at the Göttingen Register of Electronic Texts in Indian Languages (GRETIL: http://gretil.sub.uni-goettingen.de/gretil.html), the format used is more cumbersome for computational analysis.

Beyond the occasional textual inconsistencies between these editions (which tend to be minor), there is no comprehensive standard for organizing the Pali Canon. To begin with, there are slight variations in which books are considered canonical. For example, the *Milindapañha* and *Peṭakopadesa* are sometimes included in the *Khuddaka Nikāya*, and sometimes not. (They are not included in this analysis.)

Furthermore, even where the contents are agreed, the structure is sometimes not. Some elements of the overall structure are canonical and universally observed. For example, the previously discussed division of the *Tipiṭaka* into three *Piṭakas* is well established, as is the division of the *Sutta Piṭaka* into five *Nikāyas* (Webb 2011; von Hinüber 2015, 8). But beyond this, the different editions do not always agree. The division of each individual *Nikāya* into separate printed volumes is a publishing convenience and is fundamentally arbitrary. Thus, for example, the PTS edition divides the *Aṅguttara Nikāya,* or Numerical Discourses of the Buddha, into five volumes; the CST4 into only four. Both the PTS and CST4 divide the *Majjhima Nikāya*, or Middle-Length Discourses, into three volumes, but make the divisions at somewhat different (though nearby) points in the text. This means the usual standard of reference, by volume and page number, can be difficult to translate between editions.[3]

Finally, and perhaps most importantly, three characteristics of the Pali language itself create computational challenges. First, most Pali words exist in numerous declensions, generally based on number, gender, and case. Second, consecutive words in Pali sentences can be combined through letter and syllable elision in complex ways, forming what can appear to be novel words. Third, Pali also makes substantial use of compounds. Taken together, this means that individual words often appear in the Canon in a vast array of different forms.

[3] This paper largely adopts the Chaṭṭha Saṅgāyana volume numbering as a natural consequence of using an electronic version of the Chaṭṭha Saṅgāyana edition of the Canon. See the section "Abbreviations" for a longer discussion of this.

For example, there are 270 variations on the word *bhikkhu* (monk)[4] if one counts all words beginning with the base/stem *bhikkh-*. The 20 most frequent such forms are shown in Table 2. These include declensions of *bhikkhu* such as *bhikkhave*[5] (vocative plural) and *bhikkhū* (plural), related words such as *bhikkhunī* (nun), and compounds such as *bhikkhusaṅghaṃ* (congregation of monks). Of these, fully 115 (about 42%) appear in the entire Canon only once.

1. *bhikkhave*	11. *bhikkhūhi*
2. *bhikkhu*	12. *bhikkhunīnaṃ*
3. *bhikkhū*	13. *bhikkhussa*
4. *bhikkhuno*	14. *bhikkhuniṃ*
5. *bhikkhūnaṃ*	15. *bhikkhusaṅghaṃ*
6. *bhikkhunā*	16. *bhikkhavo*
7. *bhikkhuniyo*	17. *bhikkhusaṅghena*
8. *bhikkhuṃ*	18. *bhikkhusaṅgho*
9. *bhikkhunī*	19. *bhikkhūti*[6]
10. *bhikkhuniyā*	20. *bhikkhunīti*[6]

*Table 2: Most frequent words based on **bhikkhu** in the Pali Canon*

[4] All English definitions in this paper are from Buddhadatta (2014) unless otherwise noted. Where Buddhadatta gives multiple definitions, I have generally taken the first few.

[5] In fact, *bhikkhave*, the plural vocative case used in direct address, is the most common form of *bhikkhu* and appears 2.6 times as often as the nominative case that one might expect to be most common. This relatively obscure declension occurs so frequently in the Canon in conjunction with this word because many of the Buddha's discourses are directed toward a group of listening monks, whom he addresses this way. (Geiger 2005)

[6] Note that *bhikkhūti* and *bhikkhunīti* are not even single words; they are *bhikkhu* and *bhikkhunī* with the quotation marker *ti* appended. This sort of issue is discussed in more detail in the section "Limitations and future work" below.

Altogether the various discourses of the *Sutta Piṭaka* contain 115,433 distinct Pali words by our count. In comparison, the KJV contains only 13,306 words in English. Thus while the *Suttas* are less than twice as long as the Bible, they contain nearly nine times as many distinct words.

Because computational text mining typically depends on comparing word frequencies across texts, having so many words, and so many with very low frequencies, can pose a challenge. The most common word, *ca* (and; then; now), appears 56,487 times; the 100th most common word, *samannāgato* (endowed with; possessed of) appears only 2,508 times. The frequencies of all 100 most common words are shown in Figure 1, demonstrating this precipitous decline. The full lexicon of the *Tipiṭaka* follows the same frequency pattern we saw in variations of the word *bhikkhu*: about 42% of all unique words in the Pali Canon also occur only once. By way of comparison, only 31% of distinct words in the KJV appear to occur just once. (The section "Future directions" below describes some possible remediations to overcome this.)

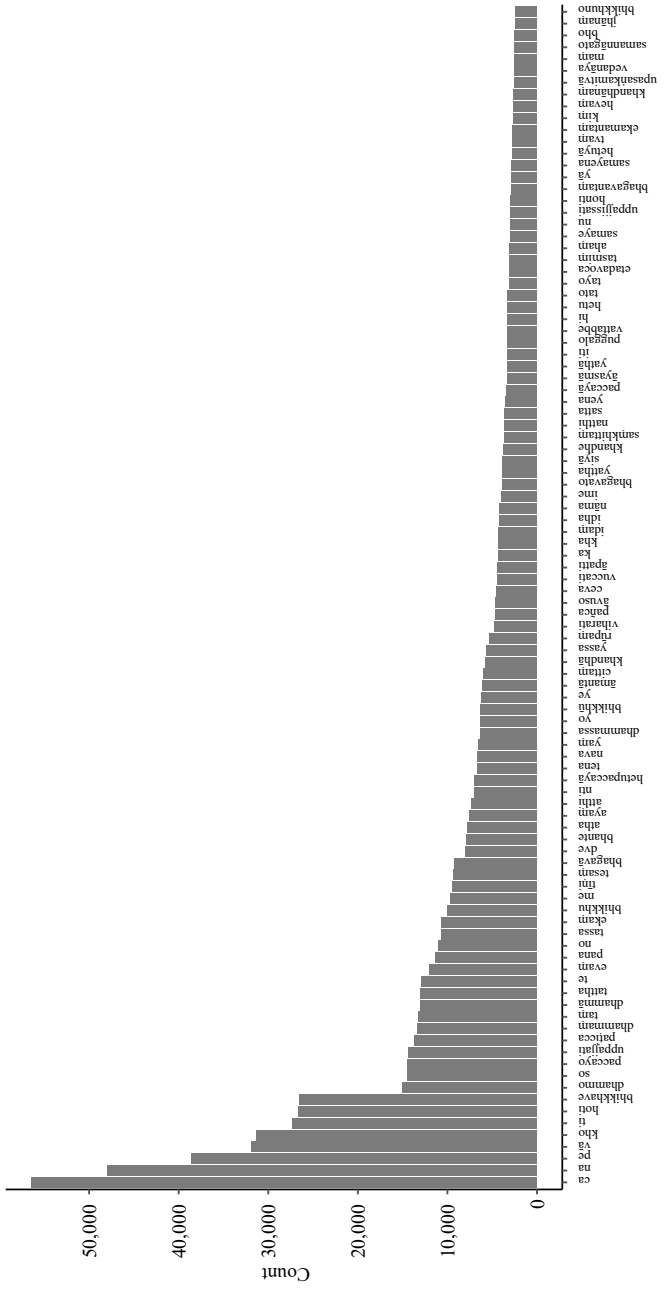

Figure 1: The 100 most common words in the Pali Canon

As shown in Figure 2, word frequency in the Pali Canon is inversely proportional to word rank. This relationship roughly follows a classic power law, as has been observed for many other language corpora (Zipf 1935). The main divergence from a Zipf power function is that the Pali Canon does not have as many high-frequency words; visually, the left side of the graph is flatter. Again, a comparison to the KJV is instructive and is shown in light gray in Figure 2. Three English words in the KJV exceed 2% frequency (*the*, *and*, and *of*), ranging from 7.5% to 4%, while no individual words in the Pali Canon are similarly common. Overall, the KJV is more "head heavy," meaning the most common words are more common than they are in the *Tipiṭaka*, while the *Tipiṭaka* itself is more "tail heavy," meaning the least common words are more common there.

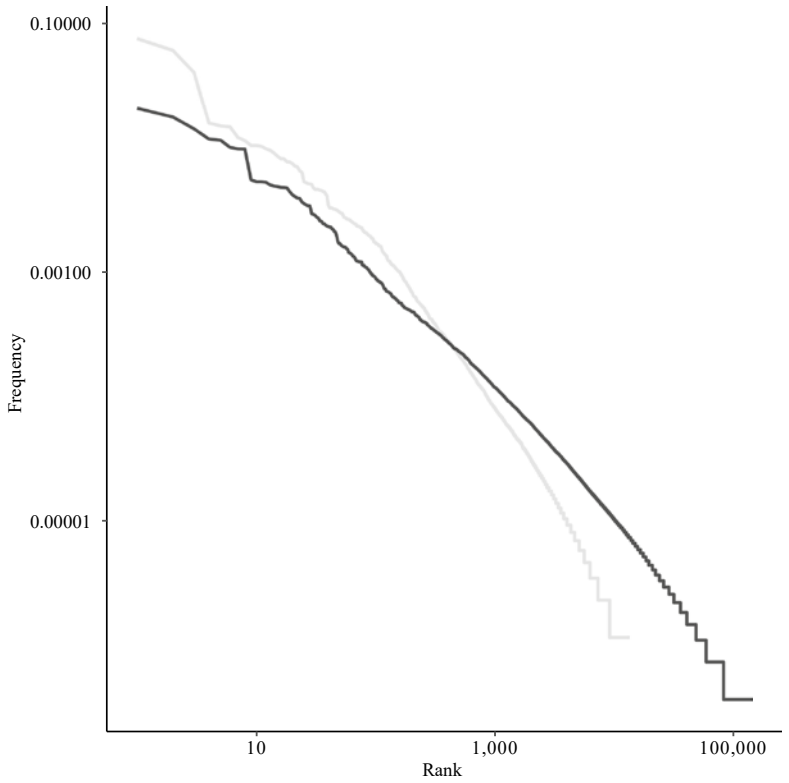

Figure 2: Word rank versus frequency across the Pali Canon (black, log-log scale), and for the King James Version of the bible (light gray)

Categorizing the *Tipiṭaka* through k-means clustering

We can further analyze the Canon using classical k-means clustering, one of the oldest algorithms for computational categorization (see MacQueen 1967 and Lloyd 1957). In the simplest terms, we compare the texts by using the relative frequency of each unique word in each text.

More precisely, this approach reduces each of our texts to some number *n* of quantitative features. The precise mechanism for transforming a text into such features is discussed below, but one can then think of these features as coordinates in an *n*-dimensional space. If our features are well chosen, then texts with coordinates closer to each other in this space should be more similar than texts further apart. In k-means clustering, we choose some number *k* of clusters in which we wish to categorize our texts, then draw boundaries in this *n*-dimensional space to create *k* distinct regions. Again, if our features are well-chosen, the points within these boundaries will form clusters of similar texts.

Of course, reducing a complex text to some manageable set of meaningful quantitative features is a difficult task. Some trivial approaches would obviously not be particularly useful. For example, the length of a text might be a poor choice because texts of the same or similar length do not necessarily have any deeper linguistic connection. Paperback editions of *The Da Vinci Code* and *A Tale of Two Cities* may both have 489 pages, but these two texts have little else in common.

A more common and often successful approach is to use the frequency of some number distinct words. In English, we might measure the relative frequency of words like *the*, *and*, *of*, *to*, and *that* (the five most common words in the KJV) as our features, such that the KJV would be represented as the five quantities {0.075, 0.060, 0.040, 0.016, 0.015}. Another text similar to the KJV would be presumed to have similar frequencies; in other words, it would occupy a nearby position if plotted in five-dimensional space. It might seem surprising that such mundane quantities could yield a useful analysis, but such analyses have led to genuinely novel discoveries in other domains of literature and the humanities (Jockers and Thalken 2020).

In this analysis of the *Tipiṭaka,* we will use the relative frequency of the 1,000 most common words in the Canon as our features. This number is admittedly arbitrary, but the results appear similar across a wide range of thresholds. By choosing the top 1,000, we are using words that appear more than 250 times across the Canon. If we were to use the top 10,000 words, we would be including

words that appear just 15 times, or less than once per volume.[7] This might lead to clusters determined by the presence or absence of a single word, or even a single typographical error in our files.

Given the discussion above on the number of distinct words in the Pali Canon, the top 1,000 may seem like a very small subset to use for our analysis; it represents well under 1% of all distinct words. However, it is well established that frequency variation among a very small number of words can often be enough to identify authorship of past literary works (Jockers & Thalken 2020). In fact, as we will discuss in our analysis of the *Sutta Piṭaka,* we can make meaningful categorizations of canonical text based on the relative frequency of far fewer than 1,000 words.

Our full methodology is roughly as follows:

1. The texts are read and separated into distinct words.
2. All numerals marking verses, pages, etc. are removed.
3. Each distinct word is counted, as well as the total words for each volume.
4. The relative frequency is computed for each distinct word (i.e., the count of that word divided by the total words in a given volume).
5. The 1,000 words with the highest average frequency across all volumes of the Canon are selected as features.
6. The distance between each volume and every other volume is calculated within this 1,000-dimensional space.
7. Boundaries are drawn to create two clusters within the space.

In effect, we are categorizing each volume of the *Tipiṭaka* based on the relative statistical distribution of the 1,000 most common words. The underlying hypothesis is that volumes with a more similar pattern of word usage are intrinsically closer (i.e., more related) to each other than those with a more dissimilar pattern of word usage.

[7] As it happens, the 1,000th most frequent word is *viññāṇassa* (a declension of *viññāṇa*: animation; consciousness). The 10,000th word is *thīnaṃ* (a declension of *thīna*: unwieldliness; impalpability).

For visual simplicity, this 1,000-dimensional space can be represented as a simple two-dimensional chart, as shown in Figure 3. These two dimensions are created by combining many of the underlying dimensions, with some loss of information, in a long-established statistical process known as principal component analysis. As shown on the axis labels of Figure 3, these two "principal components" capture approximately 84.5% of the variation between our texts in the full 1,000-dimensional space. (Volumes of the *Tipiṭaka* are shown in all figures using the standard abbreviations from the PTS Pali-English Dictionary, which are fully explained in Table 3 in the "Abbreviations" section below.)

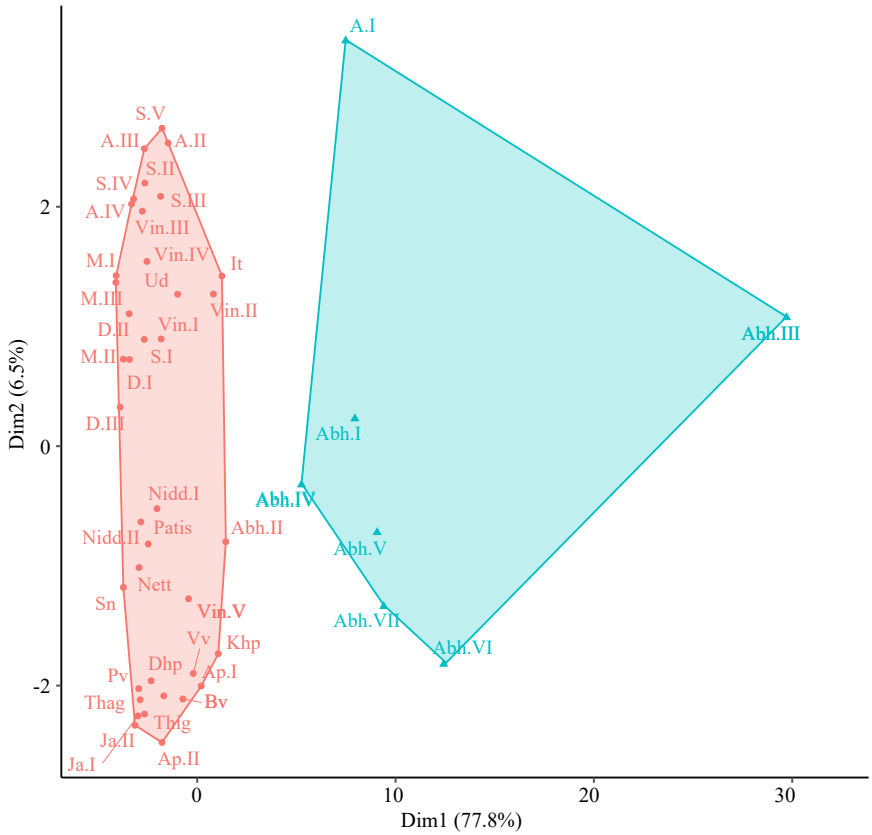

Figure 3: Cluster analysis of the ***Tipiṭaka***

We can see that the volumes of the *Tipiṭaka* shown here form two distinct clusters. The first of these contains the *Vinaya* and *Suttas* (with one exception; more on this below) while the second contains the *Abhidhamma* (again with one exception). This division roughly follows scholarly opinion on the age of the material; the Abhidhamma is considered the most recent of the three baskets of the *Tipiṭaka* (von Hinüber 2015, 64). Thus we might consider the left (blue) cluster to be our older texts and the right (red) cluster to be our younger texts.

How then to explain the two exceptions to this otherwise clean separation of our texts into older and younger clusters? The second volume of the *Abhidhamma*, titled the *Vibhaṅga* and shown as Abh.II in our figure, is clustered on the left, with our older texts, although the *Abhidhamma* is generally believed to be younger. However, the *Vibhaṅga* is believed likely to be the oldest of the *Abhidhamma* material, with some dating it to a similar period as the *Vinaya* and *Suttas* (von Hinüber 2015, 69). It is thus not entirely surprising that our algorithms might place it with the older material, which it likely matches in linguistic style. Also note that Abh.II is about equidistant from its nearest volume of the *Abhidhamma* (Abh.IV) as from the nearest volumes of the *Suttas* (Patis and Nidd.I) and *Vinaya* (Vin.V). It may represent an intermediary between these two periods of scripture.

This leaves the first volume of the *Aṅguttara Nikāya*, A.I in our figure, which is shown in right/younger (red) cluster, despite being a volume of the original *Suttas*.

The *Aṅguttara Nikāya,* or "Numerical Discourses" (Bodhi 2012), is an unusual collection. The volumes are organized according to number so that we have the "book of ones," "the book of twos," etc. (von Hinüber 2015, 76). The first volume, the *Ekakanipāta*, is the book of ones, containing discourses referring to a single thing. For example:

> Bhikkhus, I do not see even one other thing that when developed leads to such great good as the mind. A developed mind leads to great good. (A I 6; Bodhi 2012)

As this example demonstrates, many of these passages are extremely short. Although there are some counter examples, most verses contain two sentences with a total of a few dozen words. In this way, the book stands somewhat apart from the other collections, and is not particularly similar to any of them – a characteristic our computational analysis correctly highlights.

In fact, the nearest neighbor of A.I in Figure 3 is A.II, which in the CST4 electronic edition contains the *Aṅguttara Nikāya* books of two, three, and four, despite the fact that A.I is clustered with the *Abhidhamma* volumes. In some sense, the grouping of A.I with the *Abhidhamma* may represent a limitation of the standard clustering algorithms, which attempt to construct compact polygons around the individual points. While A.I is closer to points in the left (blue) cluster and one can imagine extending that cluster to include A.I, the resulting polygon would be less compact, because A.I is somewhat further from the center of the left polygon than from the center of the right.

Note that Abh.III represents a significant outlier from even the other volumes of the *Abhidhamma.* (In fact, if we divide our volumes into three clusters, our algorithm places Abh.III in a cluster of its own.) One explanation is that the Abh.III, or the *Dhātukathā*, may be younger than the preceding volumes, and appears not to have been recited at the first three Buddhist Councils at all (von Hinüber 2015, 69).

Figure 4 provides another view of these "distance" measures in a hierarchical manner, using a cluster dendrogram to visualize the similarities and dissimilarities among *Tipiṭaka* texts (Kassambara 2017). The y-axis represents distance between the texts, so texts that are joined higher are less similar than those joined lower. Color coding is used to cluster these texts into distance groups. The seven "rainbow" texts on the left are all quite distant from the rest; as in Figure 3, these include most of the *Abhidhamma* (with the exception of Abh.II) as well as the first volume of the *Aṅguttara* (A.I). The remaining texts of the *Sutta* and *Vinaya Piṭaka* form two broad clusters. On the far right, we have most texts of the *Khuddaka Nikāya* (plus Abh.II); in the middle we have the first four *Nikāyas* of the *Sutta Piṭaka* as well as the *Vinaya Piṭaka.* Once again, this largely seems to reflect the scholarly consensus concerning the age of the underlying texts, as will be discussed further in the next section.

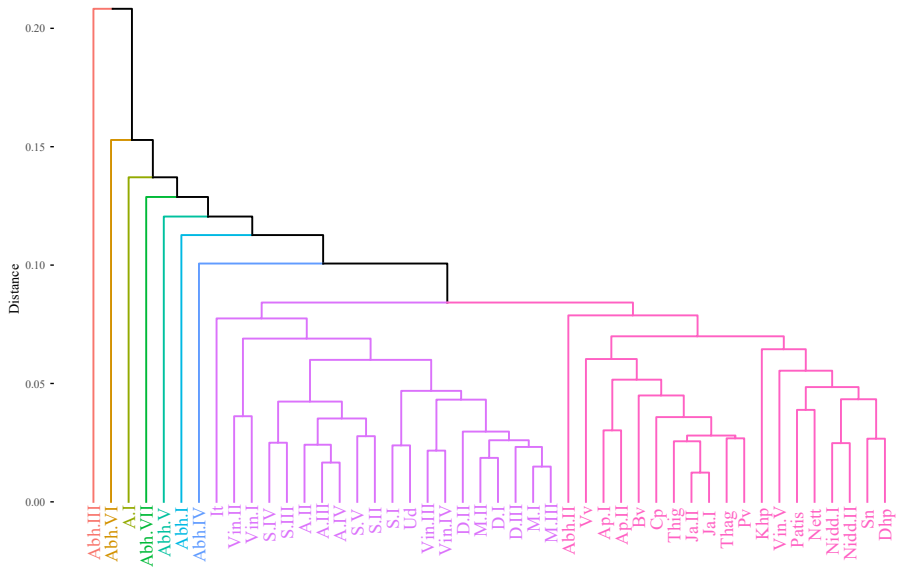

*Figure 4: Distance (i.e., dissimilarity) between **Tipiṭaka** texts*

Categorizing volumes of the *Sutta Piṭaka*

If we confine our attention to the *Sutta Piṭaka*, we can apply the same techniques and further divide these volumes into two more clusters, shown in Figure 5 below.

The upper (blue) cluster contains the *Dīgha, Majjhima, Saṃyutta*, and *Aṅguttara Nikāyas*, while the lower (red) cluster contains the many volumes of the *Khuddaka Nikāya* (labeled according to their individual volume names, as is customary) – with two exceptions. The *Udāna* (Ud) and *Itivuttaka* (It) are clustered with the volumes of the other *Nikāyas* instead of the *Khuddaka Nikāya*, where they are canonically placed. (More on this below.)

Here the clustering does not quite so closely mirror the scholarly consensus on the age of the underlying material. The *Khuddaka Nikāya* or "minor texts" represents something of a hodgepodge of "very heterogenous works" (von Hinüber 2015, 41) that appear to have been collected later than the other *Nikāyas*.

While some of these, such as the *Dhammapada* (Dhp) quoted earlier, are well-known and well-loved among Buddhists, they are generally quite distinct from the other *Sutta* collections. Von Hinüber (2015, 45) goes so far as to say that many Dhp verses "have hardly any relation to Buddhism." It therefore seems sensible that these texts can be linguistically distinguished from the first four *Nikāyas*.

While the *Udāna* (Ud) and *Itivuttaka* (It) are among the oldest elements of the *Khuddaka Nikāya*, others of likely similar age include the previously mentioned Dhp and the *Suttanipāta* (Sn). This suggests our algorithm is not clustering these texts by age per se, or at least not by age alone. On the other hand, Ud and It generally take the form of *suttas* or discourses, whereas many of the other *Khuddaka Nikāya* texts do not. In some cases, material from Ud and It is also found elsewhere in the Canon, creating inherent similarities. The placing of these texts in the upper (older) cluster may result from these textual and stylistic elements rather than, or in addition to, age.

Sn may create particular challenges for algorithms of this sort, because it is itself a collection of diverse texts of varying ages (Norman 2010, xxxi–xxxiii). The proximity of Sn to the *Niddesas* (Nidd.I and Nidd.II) in Figure 5 is likely due to the latter texts being commentaries on sections of the former. Nidd.I and Nidd.II are clearly much later than Sn, so the relationship uncovered here is not chronological but perhaps simple concordance.

Note that once again, A.I is very much an outlier with respect to the other *Suttas*. Although it falls within the upper (older) cluster, it is not particularly close to any other volume there. It is closest to It, which is also organized numerically. In fact, if we group into three clusters using the same algorithm, A.I ends up in a separate cluster of its own.

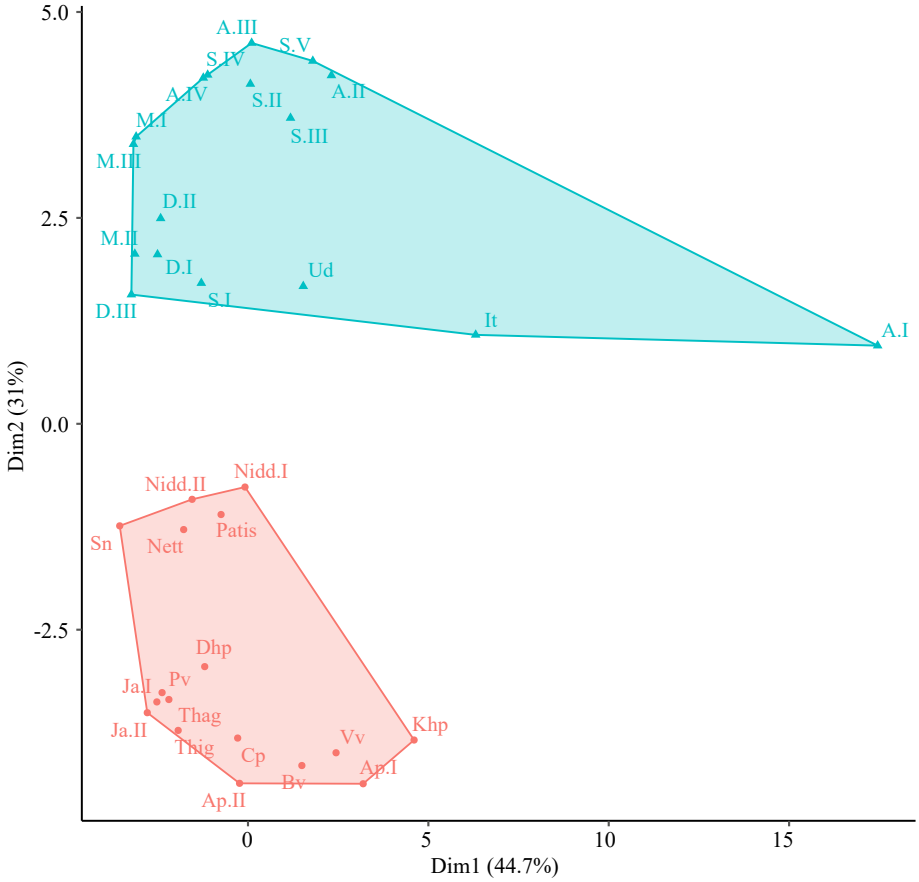

Figure 5: Cluster analysis of the ***Sutta Piṭaka***

As in the analysis of the full Tipiṭaka above, this clustering was based on the relative frequency of the 1,000 most common words in the canon. As it turns out, the suttas can be similarly categorized using a much smaller set.

Figure 6 shows a clustering of the *Sutta Piṭaka* based only on the 13 most common Pali words, which represent all the words with an average frequency of at least 0.5% across the Canon.[8] Although the shape of the clusters is inevitably different, the results are exactly the same as the 1,000-word clustering. We are able to distinguish the predominantly older and younger *suttas* based only on their use of the following words: *ca, na, kho, vā, ti, bhikkhave, hoti, pe, te, so, dhammā, taṃ,* and *me*, most of which are simple grammatical particles and the like.[9]

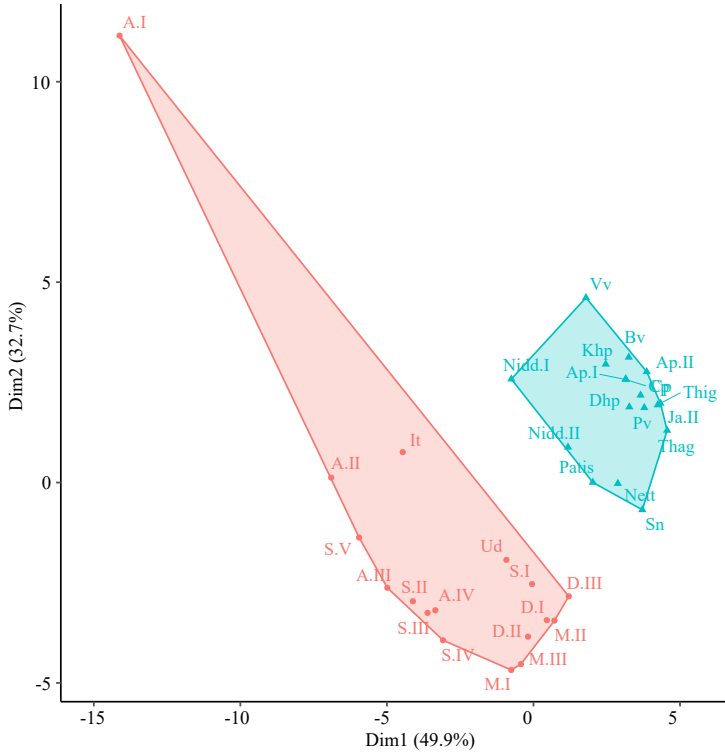

*Figure 6: Cluster analysis of the **Sutta Piṭaka** based on the top 13 words*

[8] Note that only 5 words have an average frequency of 1%, a further testament to the great linguistic variety of the canon. In contract, the KJV has 11 words with a frequency of at least 1%, and 29 with a frequency of at least 0.5%.

[9] As in the previous two analyses, A.I is a substantial outlier; in fact, even more distant from all other texts. This remains somewhat of a mystery.

We can again view this clustering hierarchically, as shown in Figure 7. We see the same broad grouping of texts into older (right) and younger (left) clusters, with A.I standing out as most distant within its cluster.

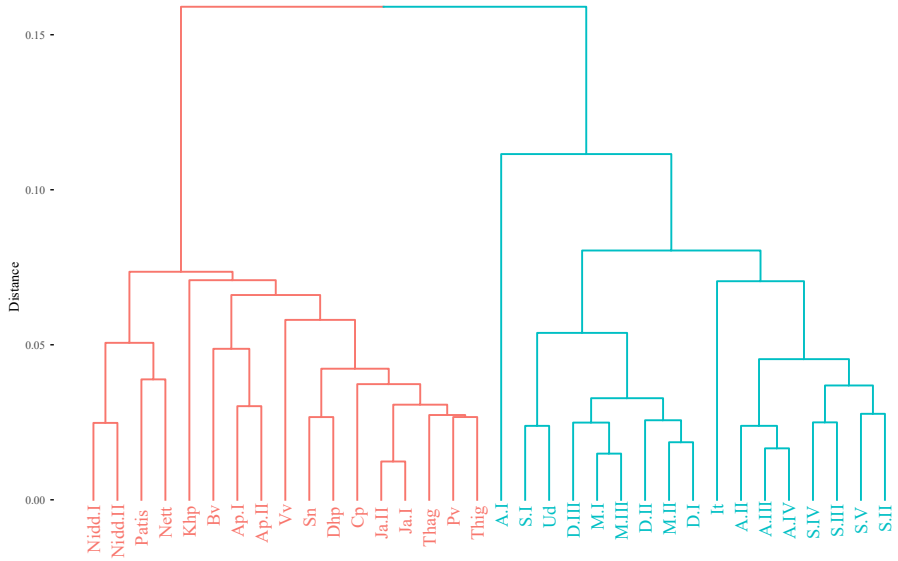

*Figure 7: Distance between **Sutta Piṭaka** texts*

Summarizing the *Sutta Piṭaka*

Word frequency can also provide clues to the core meaning of different volumes of the Canon. A common way to illustrate this is with "word clouds," graphical arrangements of individual words where the size of each word is in proportion to its frequency. In order to focus on the words with the most semantic content, a set of very common "stop words" are first removed from the corpus (Lewis et al. 2004). For example, in English, words such as *a*, *and*, and *the* are prototypical stop words: common in virtually every text and so not a very useful guide to meaning.[10]

[10] Stop words are not removed prior to the earlier clustering analysis because the relative

As far as we know, no definitive set of stop words has been defined for Pali, so a tentative set was created for this analysis. This was derived by combining the words labeled as "indeclinable" or "participle" in the PTS Pali-English Dictionary (PTS 1925)[11] plus the most common Pali pronouns (Geiger 2005, 98–109). The full list of 245 words included is shown in Table 4 in the Appendix.

Figure 8 shows such a word cloud for the *Therīgāthā,* or Poems of the Early Buddhist Nuns (left), and the *Theragāthā*, or Poems of the Elder Monks (right). As might be expected, the most prominent word in the *Therīgāthā* is *therī* (senior nun), while the most prominent word in the *Theragāthā* is *thero* (senior monk).

Figure 8: Word cloud for the ***Therīgāthā*** *(left) and* ***Theragāthā*** *(right)*

frequency of such words can be very useful in dating and identifying authorship. However, these differences are too subtle to show up in word cloud images and tell us little about meaning. By way of example, my own use of words like *a*, *and*, and *the* might help establish that I am the author of this paper or perhaps even assist in dating this paper based on the prevailing usage of such terms, but would reveal very little else about the paper's content.

[11] These were collected using the online interface to the *PED* available through the University of Chicago's Digital Dictionaries of South Asia project, at https://dsalsrv04.uchicago.edu/dictionaries/pali/.

These word clouds are potentially more interesting when applied to smaller sections of the Canon, which are likely to be more focused in meaning. Figure 9 shows such a cloud for the *Mahāsatipaṭṭhāna Sutta* (M I 56) on the left, the '(Great) Discourse on Mindfulness Meditation' (again with stop words removed). Here we see words like *pajānātiī* (knows clearly), *viharati* (lives; abides), and *loke* (declension of *loka*, the world) emphasized, which are central to the meaning of the *sutta*. On the right we see a word cloud for the full first volume *Aṅguttara Nikāya* (A I), which covers a wide-ranging set of themes. The only substantive word that stands out is *bhikkhave*, indicating that these disparate discourses were addressed to monks but had no other obvious common thread.

Figure 9: Word cloud for the ***Mahāsatipaṭṭhāna Sutta*** *(left) and* ***Aṅguttara Nikāya I*** *(right)*

Limitations and future directions

There are advantages and disadvantages to using the exact Pali syntax found in the Canon as the basis for our analysis. By way of analogy, the English words *monk* and *monks* are obviously distinct, and different authors may vary in the relative frequency of each. On the other hand, something is clearly lost if we treat the two as entirely unrelated, with no more connection than that between *monk* and *mouse.* Yet that is exactly what we are doing when we treat *bhikkhu* and *bhikkhū* as entirely distinct words.

One approach to overcoming this limitation would be to convert words to their Pali bases or stems (in this case, *bhikkh*). One could then use the base/stem frequencies as features, either replacing or augmenting the exact word frequencies. This would also avoid the issue noted above, where, for example, *bhikkhūti* is treated as a single word when it is, in fact, a concatenation of the two words *bhikkhu* and *ti*. However, developing an accurate stemming algorithm will be a substantial undertaking. Some progress has been made by others (see, for example, Basapur 2019, Elwert 2015, and Alfter 2014), but no complete algorithm appears yet publicly available. This is important work to undertake.

Our tentative list of stop words is also unsatisfactory. It was created in a somewhat manual process that may have included errors. It is also possible that, for example, all adverbs should be added to this list. Another approach would be to add all very common words, regardless of grammatical function, although this would result in meaningful words like *bhikkhave* and *dhammo* (doctrine; nature; truth) being excluded. It will likely take a good deal of trial and error, as well as a healthy dose of human judgment, to arrive at a definitive set. Our initial list is at best a good starting place for a much longer effort.[12]

The analysis described thus far has been at the "macro" scale of entire volumes. While this is interesting, it is also limiting, and in some cases arbitrary. In the future we would like to descend to the more "micro" levels of individual *suttas* and verses. The tools we have now discard demarcation of particular verses as well as word ordering within the volumes. In order to facilitate microanalysis at the verse and word level, all such material would need to be preserved.

More advanced techniques of text mining and natural-language processing could also be applied. Topic modeling—a machine learning technique for clustering and summarizing texts—is one broad example; extraction of n-grams,

[12] Elwert (2015) alludes to a set of stop words used in that work, but it does not appear to have been published.

or key phrases, is another. However, this would depend on the sort of further advances in stemming discussed above to be truly useful.

Finally, several purely technical challenges remain. The inconsistent volume numbering between the CST4 and PTS editions is an annoyance, and the solution arrived at here, sitting in between the two, is a poor one. In the future we will edit the underlying files to match the PTS numbering for consistency with other scholarly material. This is a laborious process of careful editing and was deemed too much to attempt right now.

Tentative conclusions

The analyses described here have been largely confirmational; they do not yet bring new knowledge to the study of the Pali Canon. While the apparent separation of volumes into groups of newer and older texts generally matches scholarly consensus, the discrepancies appear to be artifacts of the algorithms rather than novel discoveries. Other analyses help us visualize the relationships between the texts and some of their central themes but are not yet revealing previously undiscovered truths.

Nevertheless this style of macroanalysis shows promising potential. As these methods are refined, they may be helpful in dating noncanonical and paracanonical texts and tracing the overall evolution of the Canon. As we expand these techniques to the level of individual *suttas* and verses, we may gain still further insight into the authorship of these various component texts.

This first analysis only scratches at the surface of these ancient scriptures, showing that modern computational tools can be applied. The tools developed have been released publicly so that other scholars may continue analysis in a similar vein. We hope this is the beginning of the application of these tools to the *Tipiṭaka* and not the end.

Abbreviations

The text and figures above generally follow the standard of the *Pali-English Dictionary* (Pali Text Society 1925) but are shown in Table 3 for clarity. Note that discrepancies between the PTS and CST4 editions make volume numbering difficult. It has been handled here (admittedly somewhat inconsistently) as follows:

- Volume numbering within the *Vinaya Piṭaka* has been adjusted to match the PTS order.[13]
- Volume numbering within the *Abhidhamma Piṭaka* is consistent between the two editions and is unchanged.
- Volume numbering within the *Dīgha Nikāya* is also consistent between the two.
- Volume division and numbering within the *Majjhima Nikāya*, *Saṃyutta Nikāya,* and *Aṅguttara Nikāya* is inconsistent and has been left according to the CST4.
- Volumes of the *Khuddaka Nikāya* are listed under their separate titles rather than by number, as is the norm for these works.

The inconsistent volume numbering for the *Majjhima Nikāya*, *Saṃyutta Nikāya,* and *Aṅguttara Nikāya* is unfortunate. Reconstructing the CST4 electronic files to follow the PTS numbering would have been possible but quite laborious and so was not attempted at this time.

Vin.I – Vin.V	*Vinaya Piṭaka* volumes I – V
D.I – D.III	*Dīgha Nikāya* volumes I – III
M.I – III	*Majjhima Nikāya* volumes I – III
S.I – V	*Saṃyutta Nikāya* volumes I – V
A.I – A.IV	*Aṅguttara Nikāya* volumes I – IV
Khp	*Khuddakapāṭha*
Dhp	*Dhammapada*
Ud	*Udāna*

[13] The CST4 numbering for what PTS labels volumes I through V would be III, IV, I, II, V.

It	*Itivuttaka*
Sn	*Suttanipāta*
Vv	*Vimānavatthu*
Pv	*Petavatthu*
Thag	*Theragāthā*
Thig	*Therīgāthā*
Ap.I	*Therāpadāna*
Ap.II	*Therīapadāna*
Bv	*Buddhavaṃsa*
Cp	*Cariyāpiṭaka*
Ja.I – J.II	*Jātaka* volumes I – II
Nidd.I	*Mahāniddesa*
Nidd.II	*Cūḷaniddesa*
Patis	*Paṭisambhidāmagga*
Nett	*Nettippakaraṇa*
Abh.I – Abh.VII	*Abhidhamma Piṭaka* volumes I – VII

*Table 3: Abbreviations for **Tipiṭaka** volumes used in figures*

Acknowledgements

The texts used in this analysis were generously provided by Frank Snow at Tipitaka.org. Boris Veytsman and Gully Burns provided valuable feedback on an early draft, as did an anonymous peer reviewer, who also corrected some of my misconceptions on the *Khuddaka Nikāya* in general and *Suttanipāta* in particular. Alexander Wynne corrected several more errors in his thorough editing of the manuscript, which I greatly appreciated. I was introduced to the techniques of computational text mining at the Chan Zuckerberg Initiative in 2019–2020 under the tutelage of the superb team of research scientists working there, including Drs. Veytsman and Burns, as well as Ana-Maria Istrate, Sunil Mohan, and Ivana Williams. This work would also not have been possible without the software packages cited in the text and the great labor that has gone into developing and maintaining them, for which I am very grateful.

Bibliography

Alfter, D. (2014). "Morphological analyzer and generator for Pali." Bachelor thesis, University of Trier, Department of Digital Humanities and Computational Linguistics. https://arxiv.org/pdf/1510.01570.pdf

Basapur, S., V, Shivani, and Nair, S. S. (2019). "Pāli Sandhi – A Computational Approach." *Proceedings of the 6th International Sanskrit Computational Linguistics Symposium.*

Bodhi, Bhikkhu (trans). (2012). *The Numerical Discourses of the Buddha*. Wisdom Publications.

Buddhadatta, A.P. (2014). *Concise Pali-English Dictionary*. Motilal Banarsidass Publishers.

Carter, J.C., and Palihawadana, M. (1987). *The Dhammapada*. Oxford University Press.

Elwert, F., Sellmer, S., Wortmann S., Pachurka, M, Knauth, J., and Alfter, D. (2015) "Toiling with the Pāli Canon." *Proceedings of the Workshop on Corpus-Based Research in the Humanities*. Warsaw, Poland.

Fellows, I. (2018). *wordcloud: Word Clouds*. R package version 2.6. https://CRAN.R-project.org/package=wordcloud.

Fronsdal, G. (2005) *The Dhammapada*. Shambhala Publications.

Geiger, Wilhelm. (2005) *A Pāli Grammar.* Pali Text Society.

Gombrich, Richard (2018) *Buddhism and Pali*. Mud Pie Books.

von Hinüber, Oskar. (2015). *A Handbook of Pali Literature*. Munshiram Manoharlal Publishers.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods & Literary History.* University of Illinois Press.

Jockers, K.L and Thalken, R. (2020). *Text Analysis with R: For Students of Literature.* Springer Nature Switzerland.

Johnson, K. P., Burns, P., Stewart, J and Cook, T. (2014). *CLTK: The Classical Language Toolkit*. https://github.com/cltk/cltk

Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. STHDA.

Kassambara, A. and Mundt, F. (2020). "factoextra: Extract and Visualize the Results of Multivariate Data Analyses." https://CRAN.R-project.org/package=factoextra

Kulkarni, M. and Dangarikar, C. (2013). *Recent Researches in Sanskrit Computational Linguistics: Fifth International Symposium IIT Mumbai, India, January 2013 Proceedings.* D.K. Print World.

Lewis, D. D., Yang, Y, Rose, T. G., and Li, F. (2004). "RCV1: A New Benchmark Collection for Text Categorization Research." *J. Mach. Learn. Res.* **5**, 361–397.

Lloyd, S. P. (1957, 1982). Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in *IEEE Transactions on Information Theory*, **28**, 128–137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds L. M. Le Cam & J. Neyman, **1**, pp. 281–297. Berkeley, CA: University of California Press.

Norman, K. R. (2001). *The Group of Discourses*. Pali Text Society.

Pali Text Society. (1925). *The Pali Text Society's Pali-English Dictionary*. Pali Text Society.

Project Gutenberg (2020). *The King James Bible*. http://www.gutenberg.org/ebooks/10900.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Silge, J., and Robinson, D. (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS*, 1(3). doi: 10.21105/joss.00037, http://dx.doi.org/10.21105/joss.00037.

Sujato, Bhikkhu, and Brahmali, Bhikkhu. (2014). *The Authenticity of the Early Buddhist Texts*. Supplement to *Journal of the Oxford Centre for Buddhist Studies*, Volume 5 (https://ocbs.org/journal-supplements/)

Vipassana Research Institute (2020). *Chaṭṭha Saṅgāyana Tipiṭaka Version 4.0.* http://tipitaka.org/

Webb, Russell (ed.). (2011). *An Analysis of the Pali Canon.* Buddhist Publication Society.

Wickham et al., (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, 4(43), 1686, https://doi.org/10.21105/joss.01686

Zigmond, D. (2020). "tipitaka: Data and Tools for Analyzing the Pali Canon." R package version 0.1.1. https://CRAN.R-project.org/package=tipitaka

Zipf, G. K. (1935) *The Psychobiology of Language*. Houghton-Mifflin.

Appendix

ati	*kati*	*tassaṃ*	*paricca*	*re*
atīva	*kadā*	*tassā*	*pariññā*	*labbhā*
atha	*kamhi*	*taṃ*	*pariyādāya*	*lesa*
atho	*kayaṃ*	*tā*	*pātur*	*va*
adu	*kasmā*	*tāni*	*pi*	*vaka*
anu	*kasmiṃ*	*tāya*	*puna*	*vata*
anti	*kassa*	*tāyaṃ*	*purā*	*vā*
anto	*kassaṃ*	*tāyo*	*pure*	*vāhasā*
api	*kassā*	*tāsaṃ*	*ba*	*vi*
abhito	*kaṃ*	*tāsānaṃ*	*byā/vyā*	*vinā*
ambho	*kā*	*tāsu*	*bha*	*vinidhāya*
amma	*kāni*	*tāhi*	*bhaṇe*	*viparakkamma*
amhākaṃ	*kāya*	*ti*	*bho*	*viya*
amhe	*kāyo*	*tu*	*maññe*	*vivicca*
amhesu	*kāsaṃ*	*tuṇhī*	*mama*	*visuṃ*
are	*kāsānaṃ*	*tumhaṃ*	*mamaṃ*	*vīsati*
alaṃ	*kāsu*	*tumhākam*	*mayā*	*ve*
alālā	*kāhi*	*tumhākaṃ*	*mayi*	*vo*
assu	*kiṇi*	*tumhe*	*mayhaṃ*	*sakkā*
aha	*kim*	*tumhesu*	*maṃ*	*samma*
ahaṃ	*kimhi*	*tumhehi*	*mā*	*sammā saha*
ahe	*kismā*	*tuyhaṃ*	*murumurā*	*sā*
aho	*kismiṃ*	*tuvaṃ*	*me*	*sāgataṃ*
ā	*kiṃ*	*te*	*yagghe*	*su*
ādu	*ke*	*tena*	*yadi*	*suṭṭhu*
āma	*kena*	*tesaṃ*	*yamhā*	*sudaṃ*
ārabbha	*kesaṃ*	*tesānaṃ*	*yamhi*	*suru*
ārā	*kesānaṃ*	*tesu*	*yasaṃ*	*sū*
āsajja	*kesu*	*tehi*	*yasānaṃ*	*so*
āsu	*kehi*	*tvayā*	*yasmā*	*soḷasa*
iti	*ko*	*tvayi*	*yassa*	*ha*
ito	*kvaṇ*	*tvaṃ*	*yassaṃ*	*hañci*
ittha	*khalu*	*dabhakkaṃ*	*yassā*	*han*
itthaṃ	*kho*	*diṭṭhā*	*yaṃ*	*handa*
ida	*ca*	*dhi*	*yā*	*hambho*

idāni	*cana*	*na*	*yāni*	*have*
idha	*ci*	*nanu*	*yāya*	*haṃ*
iṅgha	*ce*	*nānā*	*yāyaṃ*	*haṃsi*
iva	*codanā*	*nāma*	*yāyo*	*hā*
iha	*jātu*	*nu*	*yāvatā*	*hi*
uda	*taggha*	*nūna*	*yāsaṃ*	*hiṅkāra*
udāhu	*tamhā*	*neva*	*yāsu*	*huṃ*
uddissa	*tamhi*	*no*	*yāhi*	*he*
uddhaṃ	*tayā*	*paññāya*	*ye*	*heṭṭhā*
upanidhāya	*tayi*	*paṭi*	*yena*	
upari	*tava*	*paṭikacca*	*yeva*	
upasagga	*tavaṃ*	*paṭṭhāya*	*yesu*	
ubbhaṃ	*tasmā*	*pati*	*yehi*	
ūhacca	*tasmiṃ*	*pada*	*yo*	
kacci	*tassa*	*pana*	*ruṇ*	

Table 4: Tentative "stop words" for Pali