Distinguishing Commentary from Canon: Experiments in Pāli Computational Linguistics

Dan Zigmond

Jikoji Zen Center
12100 Skyline Blvd
Los Gatos CA 95033
U.S.A.

Apple
1 Infinite Loop
Cupertino CA 95014
U.S.A.

djz@shmonk.com

Abstract

The *Tipiṭaka* or Pāli Canon is the canonical scripture of Theravāda Buddhists worldwide and is said to record the direct teachings of the historical Buddha. These texts were transmitted orally for several centuries before being recorded in written form in what is now Sri Lanka, likely around 100 BCE, in the Pāli language, a Middle Indo-Aryan dialect. A strong commentarial tradition evolved in the following centuries setting forth the orthodox interpretation of these texts, generally also written in Pāli. The oldest of these commentaries are now considered quasi-canonical themselves.

This paper explores the application of modern computational linguistics to these Pāli texts. We show that relatively simple analysis of word frequency allows us to distinguish canonical works from commentary. This builds on earlier analysis showing that the canonical texts themselves could be clustered using computational techniques to separate older and newer volumes. The success of these initial analyses suggests Pāli computational linguistics will be a fruitful area for future research.

1 Introduction

The *Tipiṭaka* or Pāli Canon records the oral teachings of the historical Buddha. They were codified in a series of "councils" in which the Buddha's followers gathered to recite his teachings orally and agree on their contents. The First Council is said to have been held almost immediately after his death around 400 BCE. The *Tipiṭaka* continued to be transmitted orally until the Fourth Council, held in what is now Sri Lanka around 100 BCE, when it was set in written form. The language used for this written edition was Pāli.

The word "Pāli" itself comes from the compound *pāli-bhāsa*, meaning "the language of the texts" (Geiger, 2005, xxiii). In other words, the language Pāli and the *Tipiṭaka* are inextricably linked. Although Pāli came to be used in other Buddhist literature, it is essentially unknown outside the Buddhist context.

Portions of the *Tipiṭaka* exist in other languages, but the Pāli form appears to be the oldest complete edition. Although the original manuscripts written on palm leaves are long-since lost, the volumes were painstakingly recopied over the centuries until the advent of automated printing. Even today, collections of hand-copied palm leaf manuscripts are preserved in many Asian countries. As Buddhism spread through Asia, the Buddhist community fragmented into a variety of sects and schools, many of which de-emphasized the *Tipiṭaka* in favor of newer scriptures.[1] However among the Therāvada communities still flourishing throughout Sri Lanka, Southeast Asia, and beyond, the Pāli Canon remains paramount.

Orthodox opinion among Theravāda Buddhists is that the *Tipiṭaka* records Buddha's literal teachings; in other words that Pāli is a representation of the spoken language of Magadha, the ancient kingdom in northern India where Buddha primarily lived and taught (Gombrich, 2018, 13). This assertion has long been viewed skeptically by scholars, particularly in the West.

[1] These schools would object to the description "newer," believing in many cases that the non-Pāli scriptures, generally recorded in Sanskrit, were rediscoveries of original teachings that had been lost.

Proceedings of the Computational Sanskrit and Digital Humanities, 18^{th} *World Sanskrit Conference 2023*, pages 213–222

However, a growing academic movement holds that the earliest portions of the *Tipiṭaka* may indeed contain something very close to the actual words the Buddha spoke (see, for example, Sujato and Brahmali (2014) and Gombrich (2018)). Although the *Tipiṭaka* has certainly undergone significant change over the centuries—for example, to demean to role of women (Anālayo, 2016)—elements of these Pāli texts may well trace back to Buddha's time and capture Buddha's words. As Gombrich (2018, 1) puts it, Pāli itself may represent "the argot of the Buddha and his earliest followers, […] the idiosyncratic language used by the Buddha as he toured northeast India." This possibility brings further urgency to the critical study of the *Tipiṭaka*, if only to determine the relative age of the various texts and to provide clues as to which may, in fact, have been "spoken by the Buddha" (Sujato and Brahmali, 2014, 7).

The name *Tipiṭaka* literally means "three baskets" and derives from the traditional division of the Canon into three distinct collections of texts:

- *Vinaya Piṭaka*, "Basket of Discipline," describing the rules for Buddhist monks and nuns, their origin, and their evolution.
- *Sutta Piṭaka*, "Basket of Teachings," compiling the oral teachings of the Buddha and a few of his most notable disciples.
- *Abhidhamma Piṭaka*, "Basket of Special Teachings" or "Basket About the Teachings," explaining and systematizing various Buddhist doctrines.

Over time Pāli became the preferred language for most Buddhist ecclesiastical writing in the Theravāda tradition.[2] A prodigious commentarial literature evolved, describing the orthodox interpretation of the *Tipiṭaka* for generations of Buddhists (von Hinüber, 1997, 100). Among the most revered of these commentaries are those believed to have been composed by the teacher Buddhaghosa, likely between 370 and 450 CE (von Hinüber, 1997, 103). Eventually, others composed commentaries on the commentaries (usually referred to as "sub-commentaries"), resulting in a vast, sprawling Pāli corpus.

Despite this extensive body of writing, Pāli has no fixed written form. It is traditionally transcribed using the native alphabets of the countries in which Therāvada Buddhism is practiced: Sinhala in Sri Lanka, Khom and Tham in Thailand, Burmese in Burma, Khmer in Cambodia, and Devanagari in India. In the West it is generally written in Roman script, using the International Alphabet of Sanskrit Transliteration long adopted for Sanskrit and other Indic languages. Thus, for example, *bhikkhu* and *bhikkhuṇī* (monk and nun) in Roman script are equivalent to भिक्खु and भिक्खुणी in Devanagari.

2 Pāli Computational Linguistics

Computational linguistics—the application of computational techniques to the study of language—is nearly as old as digital computing itself. The recent explosion in the availability of both computing power and electronic texts has caused a concomitant explosion of research in this field. Although much of this work was initially in military, governmental, and commercial contexts, recent decades have seen a flourishing of applications in the humanities (see, for example, Jockers (2013) and Jockers and Thalken (2020)). There has also been considerable success using computational techniques to solve the specific problem of determining the authorship of texts (see, for example, Rosen-Zvi et al. (2010)).

Historically much of this work was heavily biased towards English and a handful of other modern languages. By one estimate, as recently as 2007, "only a very small number (perhaps thirty) of the world's 6000+ languages currently enjoy[ed] the benefits of modern language technologies" (Scannell, 2007). Even among modern European languages, the Multilingual Europe

[2] As Buddhism spread to other countries within Asia, Pāli was also used as a means of communication between monks and nuns who did not share a native language. This practice seems to have died out, in part due to the spread of English as a second language. The last book on spoken Pāli appears to be have been published in 1951 (Buddhadatta, 1951) and is long out of print.

Technology Alliance found in 2013 that only English had consistently "good support" across technology categories, and only French and Spanish had consistently "moderate support" across all categories (Multilingual Europe Technology Alliance, 2013).

Fortunately this has started to change, with under-resourced languages receiving increasing attention in recent years. In particular, efforts to apply computational techniques to ancient languages is picking up speed. The Classical Language Toolkit provides resources for a wide variety of ancient languages (Johnson et al., 2014). Recurring international symposia have been held on Sanskrit computation linguistics since 2007 (Gerard, 2009).

Yet relatively little computational research has been applied to Pāli thus far. Elwert et al. (2015) began work on a structured electronic edition of the *Sutta Piṭaka* in Pāli, but this work was never completed and the corpus never published. Alfter (2015) developed several tools for analyzing Pāli texts in Java, but that work appears to be no longer maintained. More recently, Haribhakta and Nadageri (2017) have explored labeling parts of speech in Pāli sentences, and Basapur et al. (2019) describe a computational approach to the issue of Pāli word splitting, which (as in Sanskrit) is complicated by Pāli's sandhi rules for word combining and elision. (Basapur et al. (2019) use Devanagari rather than Roman script, which may make the work less accessible to some Pāli scholars.) Beyond this handful of papers, little else has been published.

Efforts are now in place to facilitate broader application of computational techniques to Pāli. Last year we published `tipitaka` (Zigmond, 2020), a package for Pāli computational linguistics using the R statistical software language (R Core Team, 2020). It is currently in a nascent stage, providing access to the *Tipiṭaka* in raw and lightly-processed form and a few basic tools for sorting and comparing Pāli words. Our first applications of this package to the analysis of the *Tipiṭaka* showed some promising results. For example, rudimentary word frequency analysis was able to separate older and newer volumes of the *Tipiṭaka* into the clusters that roughly matched the scholarly consensus on their relative age (Zigmond, 2021).

3 Distinguishing Commentary from Canon

The results described here extend that early work to include the Pāli commentary as well as the Canon. Here we focus exclusively on the *Sutta Piṭaka*, the most widely read and studied of the three baskets of the *Tipiṭaka*. As in our prior study, we divide our texts into two groups using k-means clustering (MacQueen, 1967; Lloyd, 1982), which can be thought of as a simple form of unsupervised machine learning. In this case, we use the algorithm to classify our texts into the two implicit categories of canon and commentary. As in Zigmond (2021), the frequencies of the global top 1,000 Pāli words (i.e., the 1,000 words most frequently found across the entire corpus) become our features, although past work suggests the results are typically robust across many frequency thresholds, both higher and lower than 1,000.[3]

The `tidy` (Wickham et al., 2019) and `tidytext` (Silge and Robinson, 2016) packages in R make the text processing tasks fairly simple. Like the `tipitaka` package, this work uses the Chattha Sangāyana Tipiṭaka version 4.0 (CST4) edition of the *Tipiṭaka* (Vipassana Research Institute, 1990). Illustrations of the resulting clusters were created with the `factoextra` package (Kassambara and Mundt, 2020).

Figure 1 shows the results of this clustering. Each file in the CST4 distribution is plotted separately, and the naming convention follows that of CST4 itself: `sttvvx`, where `s` denotes the *Sutta Piṭaka*, `tt` is a two-digit text number, `vv` is a two-digit volume number, and the final letter `x` is the character `m` for canonical or root texts (from the Pāli word *mūla*, root) or `a` for commentary (from *aṭṭhakathā*, explanation or commentary). Thus, for example, `s0502m` denotes the canonical fifth text (i.e., the *Khuddaka Nikāya*), second volume (the *Dhammapada*), while `s0502a` denotes the commentary on that same volume. Where an additional numeral

[3]It's important to note that that we use the term "word" loosely here to mean any string of letters delimited by white space, punctuation, or numerals. These strings can, in practice, represent multiple Pāli words due to sandhi. While this is not quite as large an issue in Pāli and Sanskrit, it nevertheless represents a limitation of this approach, as discussed both below in Section 4 and in our prior paper (Zigmond, 2021).

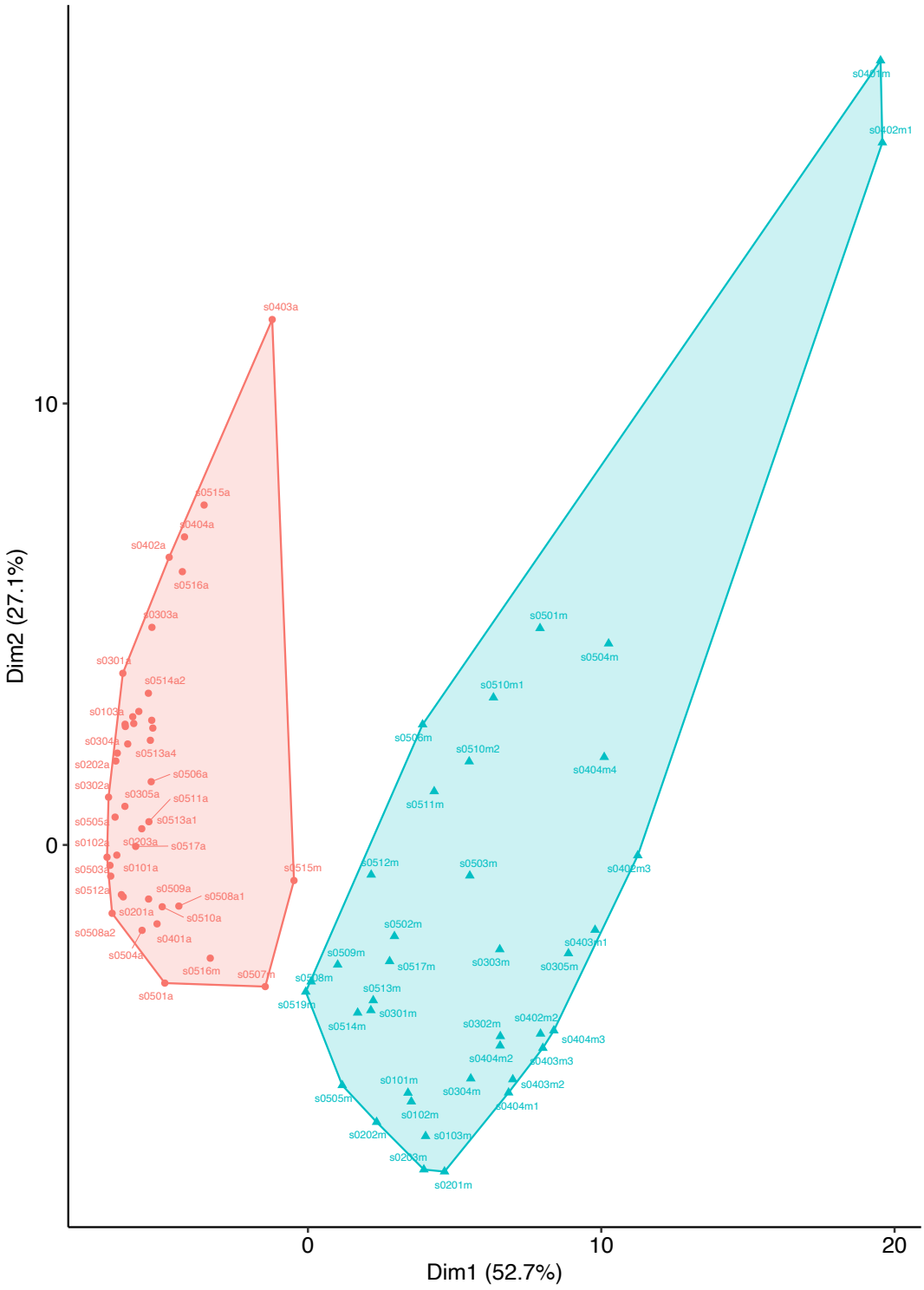

Figure 1: Clustering canon and commentary. On the right (blue) we see the canonical texts and on the left (red) we see *primarily* the commentarial texts, with three canonical texts (`s0507m`, `s0515m`, and `s0516m`, all near the bottom) miscategorized there. For naming conventions, see Table 1.

appears after this sequence, as in s0513a1, it simply means a single printed volume was split into multiple text files in the electronic edition of the CST4. A full list of CST4 file names and their corresponding *Tipiṭaka* text is given in Table 1.

We can see in Figure 1 that the two clusters separate the canonical texts from the commentary nearly perfectly, with the right cluster containing the root texts and the left cluster containing the commentary. There are three exceptions, where root texts appear to be incorrectly categorized: s0507m, s0515m, and s0516m. All of three come from the *Khuddaka Nikāya*, "the collection of short pieces" (Geiger, 1956, 19), a diverse set of texts generally considered the last part of the *Sutta Piṭaka* to be closed.

The latter two files, s0515m and s0516m, contain the *Mahāniddesa* and *Cūḷaniddesa*, respectively. These are an interesting case in that they are, in fact, commentaries on portions of the *Suttanipāta*. They are not attributed to Buddha but to Sāriputra, one of his chief disciples. They are believed to be much more recent than the original sutras, with estimates of their origin ranging from roughly 100 BCE to 200 CE (von Hinüber, 1997, 59). This is older than most of the other commentaries, but younger than most of the canon. It seems reasonable that such canonical texts could be "mistaken" for commentarial literature, since they are, in fact, both.

On the other hand, the characterization of s0507m file is more puzzling. This is another text of the *Khuddaka Nikāya*, the *Petavatthu*, a book containing "the stories of the departed ones (*petas*) who are suffering because of the bad actions they have committed during their previous existence" (Norman, 1983, 71). It is clearly also a late addition to the canon, likely added after the Second Council (von Hinüber, 1997, 51), perhaps "a short time before the third council" (Geiger, 1956, 20) or roughly 200 years after Buddha's passing (Norman, 1983, 71). Yet other texts in the *Khuddaka Nikāya* are of a similar age, including the *Vimānavatthu* (s0506m), which are not clustered with the commentaries by our algorithm. Age therefore cannot be the sole cause of the miscategorization.

It seems that this area of the clustering is somewhat unstable. The Burmese Theravāda lineages have traditionally included two additional volumes in the *Khuddaka Nikāya*: the *Milindapañha* and *Peṭakopadesa*, which are included in the CST4 as s0518m and s0520m, respectively, although the files are tagged as "miscellaneous" rather than canonical.[4] Because there is disagreement among Theravada traditions about the status of these texts, von Hinüber (1997, 76) calls them "paracanonical." If we add these to our corpus, our algorithm produces the clusters shown in Figure 2. Now only the *Cūḷaniddesa* (s0516m) is clustered "incorrectly," with the commentaries rather than root texts. Again, given that the *Cūḷaniddesa* and *Mahāniddesa* are both canonical and commentarial, it makes sense that their categorization is somewhat unstable between the two clusters, easily perturbed by other changes to the corpus of texts.

As mentioned above, eventually new commentaries were written to explain the older commentaries, referred to sub-commentaries or *ṭīkā*. Figure 3 shows the effect of adding the 16 *Sutta Piṭaka ṭīkā* included in the CST4 to the 40 *aṭṭhakathā*, 39 *mūla*, and 2 paracanonical files. (All sub-commentary file names end in t, as in sttvvt.) Here again, only the *Cūḷaniddesa* seems to be "miscategorized." All the volumes of sub-commentary are correctly clustered with the original commentaries.

4 Limitations and Conclusions

This work shows that simple k-means clustering based on unique word frequencies can reliably distinguish most Pāli canonical literature from the Pāli commentaries. The few errors we see are on the margins of the canon, with a few late additions to the *Tipiṭaka* clustered with the later commentaries. This bodes well for future applications of computational linguistics analyzing the age and authorship of Pāli texts.

That said, and as discussed in both Elwert et al. (2015) and Zigmond (2021), the Pāli Canon

[4]The *Nettippakaraṇa* (s0519m) is similarly paracanonical, but is not distinguished from other root texts in the CST4.

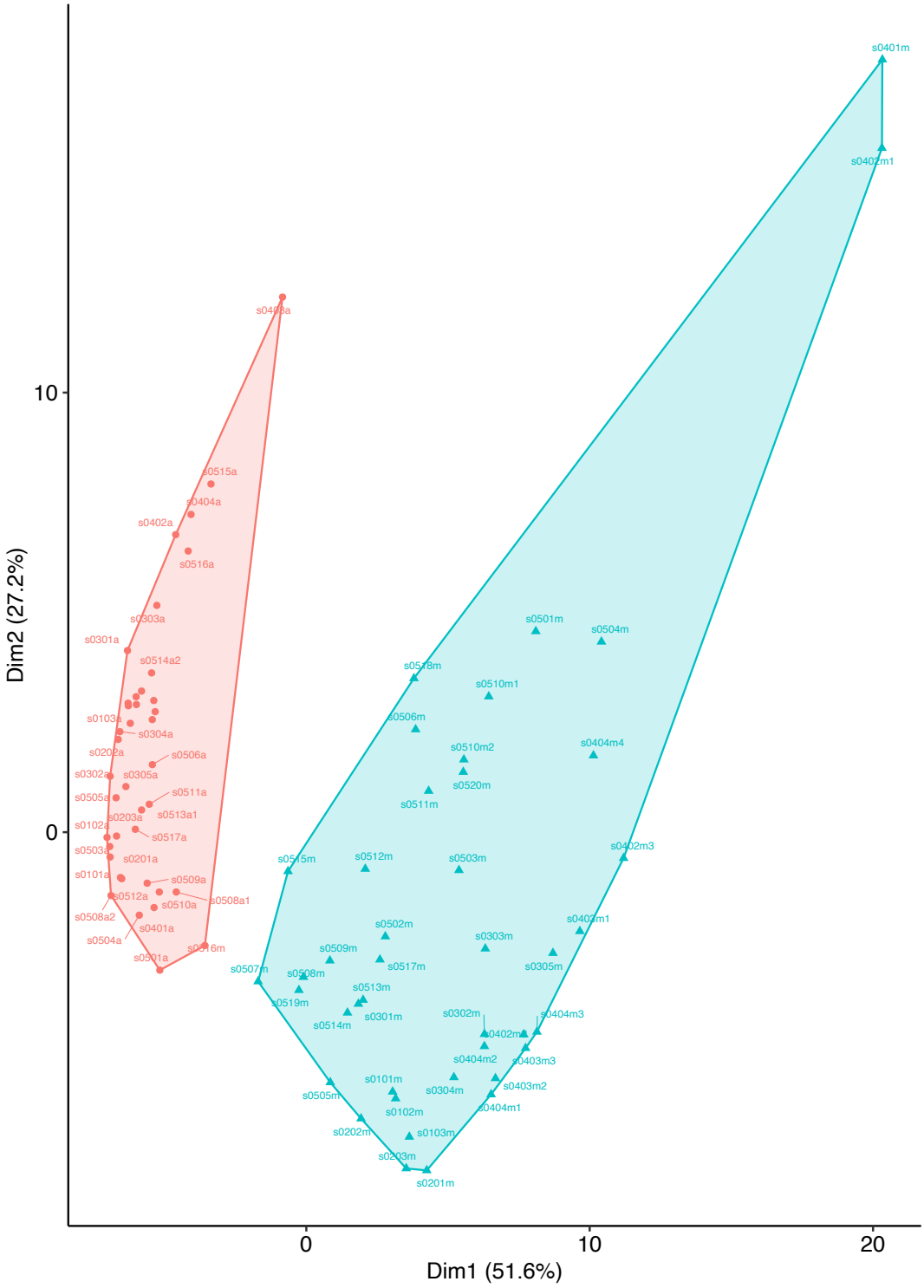

Figure 2: Adding two paracanonical texts (`s0518m` and `s0520m`) to our clustering. Again on the right (blue) we see the canonical texts and on the left (red) we see mostly commentarial texts, now with only one canonical text (`s0516m`) miscategorized.

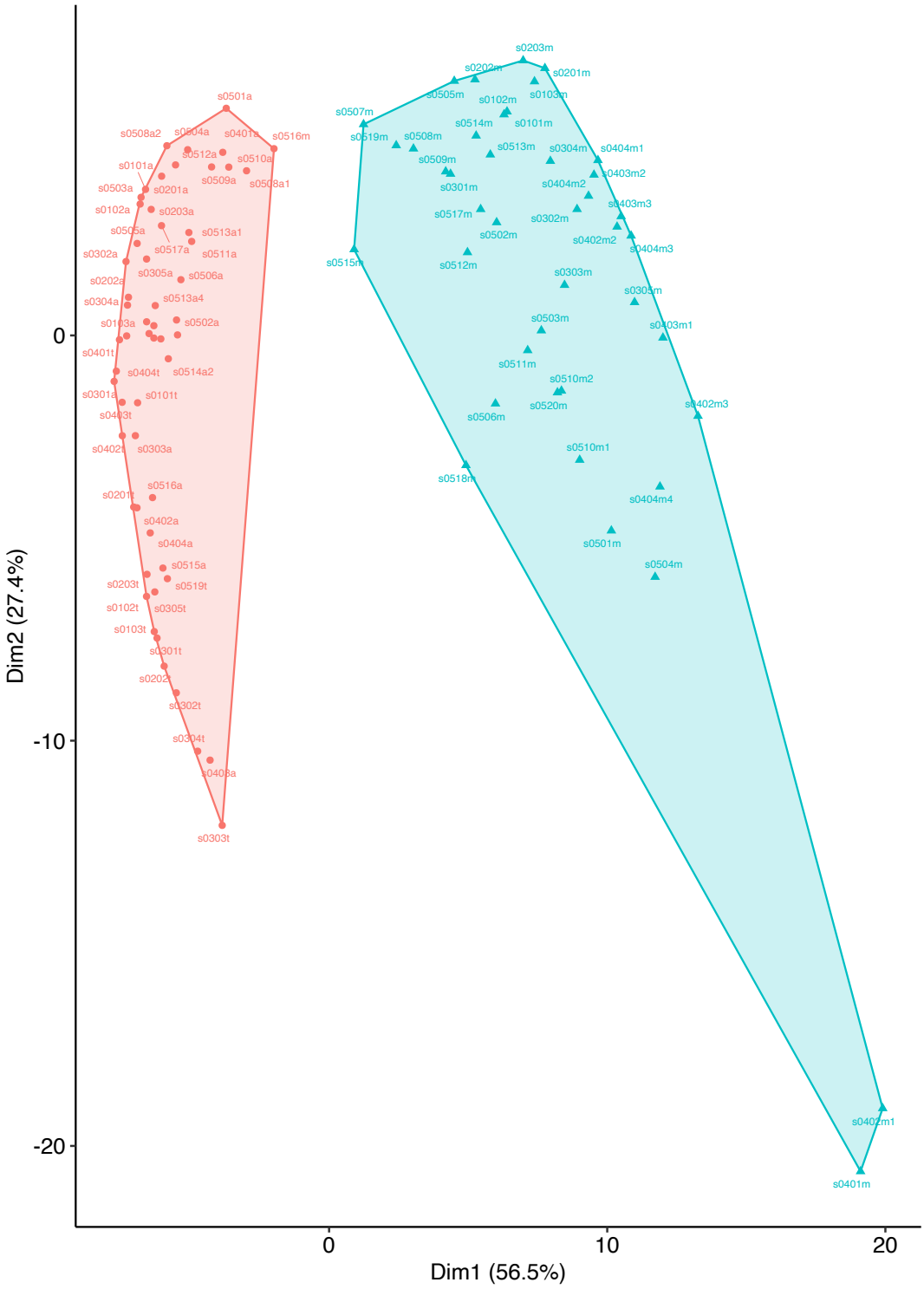

Figure 3: Adding the sub-commentaries (which follow the form to our clustering. Although the shape is different, the results are unchanged, with again only one canonical text (`s0516m`, now near the top) miscategorized.

File	*Tipiṭaka* text
s01vvx	*Dīgha Nikāya*
s02vvx	*Majjhima Nikāya*
s03vvx	*Saṃyutta Nikāya*
s04vvx	*Aṅguttara Nikāya*
s05vvx	*Khuddaka Nikāya*
s0501x	*Khuddakapāṭha*
s0502x	*Dhammapada*
s0503x	*Udāna*
s0504x	*Itivuttaka*
s0505x	*Suttanipāta*
s0506x	*Vimānavatthu*
s0507x	*Petavatthu*
s0508x	*Theragāthā*
s0509x	*Therīgāthā*
s0510x	*Apadāna*
s0511x	*Buddhavaṃsa*
s0512x	*Cariyāpiṭaka*
s0513x	*Jātaka* I
s0514x	*Jātaka* II
s0515x	*Mahāniddesa*
s0516x	*Cūḷaniddesa*
s0517x	*Paṭisambhidāmagga*
s0518x	*Milindapañha**
s0519x	*Nettippakaraṇa*†
s0520x	*Peṭakopadesa**

Table 1: File naming conventions in the CST4. Note that the individual volumes of the *Khuddaka Nikāya* are typically referenced by name, while the volumes of the first four *Nikāyas* are not. Also, * designates paracanonical volumes differentiated in the CST4, while † designates a paracanonical volume included as canonical in the CST4. The character `x` is always `m` for root (canonical) texts, `a` for commentaries, and `t` for sub-commentaries. Any number after the final letter (as in `s0403m1`) simply means the volume was split into multiple files due to length.

in raw form is a poor foundation for this sort of textual analysis. Similar words appear in a wide array of dissimilar forms, due to declensions, compounds, and sandhi. In addition, volumes are divided arbitrarily, extraneous words are sometimes added, and at times typographical errors have clearly crept in. Some of these artifacts may provide clues to the age of the texts—if, for example, different errors have appeared in different periods, or where word combining practices have evolved over time—but many of them simply add noise.

Although the present work shows that useful research can be carried out directly on the electronic edition of the CST4, Pāli computational linguistics is in dire need of a more refined corpus. At the very least, we would like to be able to run the same analysis on both raw and "corrected" or "normalized" versions of the texts. The very light pre-processing of the `tipitaka` package will not be sufficient in the long run.

Recently the Digital Pāli Tools (2021) project has embarked on an ambitious effort to create a more suitable corpus for Pāli computational linguistics and related applications. Once this work is complete, it should be possible to apply the tools and techniques used here against their new preprocessed texts. That will allow us to establish whether some of the anomalies in our results—such as the strange clustering of the *Petavatthu*—can be explained by artifacts in the raw text files. It should be relatively straightforward to adapt the `tipitaka` package to this corpus once it is available.

Many interesting analyses will be possible using these techniques against a more robust corpus. For example, we can begin looking below the level of full texts to compare verse passages to prose. We can use more advanced machine learning algorithms to attempt to classify individual chapters of heterogeneous works like the *Suttanipāta*, to explore when they may have been composed. We may be able to estimate more precisely when texts were added to the canon, taking us one step closer to determining which may be the actual words of the historical Buddha.

Acknowledgements

The files used in this analysis from the electronic edition of the CST4 were generously provided by Frank Snow of Tipitaka.org. This work was greatly facilitated by the software tools referenced earlier, including the `R` statistical programming language and the `tidy`, `tidytext`, and `factoextra` packages. I am also grateful to the four anonymous reviewers for the Computational Sanskrit and Digital Humanities section of the World Sanskrit Conference 2021–2023, who provided valuable feedback and several corrections.

References

David Alfter. 2015. Analyzer and generator for Pali.

Anālayo. 2016. *The foundation history of the nuns' order*. Projekt Verlag.

Swati Basapur, Shivani V, and Sivaja Nair. 2019. Pāli sandhi – a computational approach. In *Proceedings of the 6th International Sanskrit Computational Linguistics Symposium*, pages 181–192, IIT Kharagpur, India, October. Association for Computational Linguistics.

A. P. Buddhadatta. 1951. *Aids to Pali conversation and translation*. P.M.W. Piyaratna.

Digital Pāli Tools. 2021. A step-by-step process towards natural language processing of Pāli. https://bitly.com/dptvision.

Frederik Elwert, Sven Sellmer, Sven Wortmann, Manuel Pachurka, and David Alfter Jürgen Knauthand. 2015. Toiling with the Pāli Canon. In *Proceedings of the Workshop on Corpus-Based Research in the HumanitiesPage*, pages 39–48, Warsaw, Poland. Institute of Computer Science, Polish Academy of Sciences.

Wilhelm Geiger. 1956. *Pali literature and language*. University of Calcutta.

Wilhelm Geiger. 2005. *A Pāli Grammar*. Pali Text Society.

Huét Gerard. 2009. *Sanskrit computational linguistics first and second international symposia, Rocquencourt, France, October 29 - 31, 2007 ; Providence, RI, USA, May 15 - 17, 2008 ; revised selected and invited papers*. Springer.

Richard Gombrich. 2018. *Buddhism and Pali*. Mud Pie.

Yashodhara Haribhakta and Laxmi Nadageri. 2017. Parts of speech tagger for Pali language. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)*, 2(4):845–853, Jul–Aug.

Matthew L. Jockers and Rosamond Thalken. 2020. *Text Analysis with R.: For Students of Literature*. Springer International Publishing.

Matthew Lee Jockers. 2013. *Macroanalysis digital methods and literary history*. University of Illinois Press.

Kyle P. Johnson, Patrick Burns, John Stewart, and Todd Cook. 2014. Cltk: The classical language toolkit. https://github.com/cltk/cltk.

Alboukadel Kassambara and Fabian Mundt, 2020. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7.

S. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

J. B. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Multilingual Europe Technology Alliance. 2013. Meta-net white paper series: Key results and cross-language comparison.

Kenneth R. Norman. 1983. *Pāli literature incl. the canon. literature in Prakrit and Sanskrit of all the Hīnayāna schools of Buddhism*. Harrassowitz.

R Core Team, 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. 2010. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, 28(1):1–38.

Kevin Scannell. 2007. The Crúbadán project: Corpus building for under-resourced languages. *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, 01.

Julia Silge and David Robinson. 2016. tidytext: Text mining and analysis using tidy data principles in R. *JOSS*, 1(3).

Bhikkhu Sujato and Bhikkhu Brahmali. 2014. *The authenticity of the early Buddhist texts*. Buddhist Publication Society.

Vipassana Research Institute. 1990. Chaṭṭha saṅgāyana tipiṭaka version 4.0.

Oskar von Hinüber. 1997. *A handbook of Pāli literature*. Munshiram Manoharlal Publ.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Dan Zigmond, 2020. *tipitaka: Data and Tools for Analyzing the Pali Canon*. R package version 0.1.1.

Dan Zigmond. 2021. Toward a computational analysis of the Pali Canon. *Journal of the Oxford Centre for Buddhist Studies*, 20.