Tibetan Buddhist English: a corpus approach to the Tibetan Buddhist genre of *shastra* within the Kagyu Shedra curriculum.

I J FRYE

PhD 2022

Tibetan Buddhist English: a corpus approach to the Tibetan Buddhist genre of *shastra* within the Kagyu Shedra curriculum.

INGRID JESSICA FRYE

A thesis submitted in partial fulfilment of the requirements of

Manchester Metropolitan University

For the degree of
Doctor of Philosophy

Department of Languages, Linguistics and Communications
Manchester Metropolitan University

2022

ACKNOWLEDGEMENTS

For the ongoing support throughout a journey that has been interspersed with numerous life-changing challenges and an indescribable manifestation of kindness and support both professionally and morally, I would like to thank the following people and teams:

My supervisory team Frances Johnson, Amy Wang and Rob Drummond for their stamina and patience to help me see this thesis through to the end, and for sharing their knowledge, experience and wisdom, for their sense of humour, their pragmatism, their ability to motivate, provide meaningful feedback, stimulate critical engagement, and for their extraordinary ability to co-supervise.

Lama Jampa Thaye and the members of the Dechen Sangha without whom this project would have never been possible.

My silent co-existers Beate Peter, Emma Greenhalgh, Erin Palmer and John Rowe, who supported me in the writing process, kept me motivated and prevented me from feeling isolated all the way until the end.

The iHSSR at Manchester Met University for providing the initial funding for this project, related CPD and networking opportunities, which enabled me to develop the research expertise required for this study, especially Jenny Rowley and Berthold Schoene.

My former colleagues from MMU and my current and former colleagues from the University of West London for their inspiration and support, especially to Fereshte Goshtasbpour, Fiona Long, “Dr Phil” Morris, Andy Harris, Fred Guery, Eljee Javier, Ben Guilbauld, Joaquin Fernandez, Laura Martin-Cisnero, Isabelle van der Schelden, Benedicte Brahic, Tobias Etzold, Jannie Roed, Antony Aleksiev, Erik Blair, Hendrik van der Sluis, Diogo Casanova, Lesley-Jane Eales-Reynolds, Joelle Fanghanel, Sara Raybould and Isabel Huet.

The MMU Graduate School administrative team for their availability and support, especially Deborah Bown, who has helped me manage the administrative struggle along the way.

The Lancaster University’s Centre for Computer Corpus Research on Language (UCREL) team and Queen Mary University writing support team for opening their doors to externals and thereby enabling me to further my research and academic writing skills – including catering for my pregnancy cravings and taking me to A&E during the PhD Bootcamp.

My former lecturers Amy Wang, Huw Bell, Laurence Coupe and Carmen Herrero for all of your inspiration and for your mentorship to help me develop the confidence to pursue this project.

My internal and external examiners Dawn Archer and Ivor Timmis for their invaluable feedback and comments to help me finalise this thesis.

My family – blood, in-law or “adopted” - for giving me perspective and keeping me grounded all along the way, especially my wonderful boy Luca for reinforcing my work-life balance and challenging me to set a positive example by not giving up, my father Horst and great auntie Dorothea for their emotional and financial support and their great restraint in asking me about my thesis completion, and my late mother who, on her deathbed, made me promise to bring my PhD to completion.

To my husband “Dr John”. Without your love, your faith in my capability, your inexhaustible patience, your selfless support, your perfection in the art of coffee making, your ability to say the right thing or provide a kick in the right place at the right time, this would not have been possible.

ABSTRACT

Against the backdrop of the argument of the incomprehensibility of Buddhist English language to non-specialist audiences due to the high frequency of Sanskrit loanwords and unexplained terminology and a general lack of data-driven, empirical research on the use of Buddhist English beyond Buddhology and translation studies, this thesis investigates the following research questions: (1) What are pervasive linguistic features of the genre shastra in Tibetan Buddhist English? (2) Based on question 1, what are the characteristics of such linguistic features? (3) What is the link between such linguistic features and their situational context of Tibetan Buddhist shastras? (4) How do the linguistic features of Tibetan Buddhist shastras compare to other written registers? Compilation and frequency-based analysis of a small specialised corpus of Tibetan Buddhist Shastras (commentaries) identified four typical linguistic features: lexical closure, low type-token ratio (TTR), frequent use of the indefinite pronoun *one* and the frequent use of Sanskrit loanwords. Analysis was carried out following Biber and Conrad's (2013) framework for register analysis, comprising situational, linguistic and functional analyses. Lexical closure properties in the corpus provided a reliability measure for the findings of the study. Together with a low frequency of personal pronouns and a high frequency of the generic pronoun *one*, they aligned with characteristics of general and academic written registers. Existing characteristics of written registers have been challenged for their disassociation of high TTR and the use of the specific pronoun *one*, which in Buddhist English were found to be features of written register, indicative of the frequent repetition of titles and headings and frequent anaphoric referencing to aid the Buddhist practice of memorisation. The high frequency of loanwords proved to align with the claim of incomprehensibility of Buddhist language for a non-specialist audience, yet the relationship between situational and linguistic analysis indicated that such shortcomings of Buddhist English are mitigated through the common Buddhist practice of textual study as part of so-called "Shedras in the West". Contributions include the provision of empirical data on the under-investigated register of Buddhist English Shastras, and to register classifications of written and academic registers. Methodological contributions were made through provision of a first-ever corpus-based study of Buddhist English, thereby testing the validity of established corpus approaches in a small specialised context. Theoretical contributions included an evaluation of Biber's multidimensional analysis framework (1988, 2007), calling for an extension of the existing frameworks to account for the deviations in the findings based on the Buddhist English register shastra. Furthermore, the study provides a template for the calculation of lexical closure as a measure for representativeness in small corpora. Additional contributions are made by illustrating the pedagogic application of corpus data in the classroom by means of sample classroom tasks.

This thesis is dedicated to the memory of my mother Ingrid Dorothea Frye.

Contents

LIST OF FIGURES

LIST OF TABLES

GLOSSARY OF BUDDHIST TERMS[1]

Bodhicitta/bodhichitta/bodhimind

The vow and commitment to attain enlightenment for all sentient beings

Buddha

The awakened one

Dharma

The collection of all teachings of the Buddha.

Discourses

One collection of teachings of the Buddha.

Kagyu School (of Tibetan Buddhism)

One of the four major lineages (see below) of Tibetan Buddhism

Lama

Spiritual teacher

Lineage

A lineage in the context of Tibetan Buddhism is the continued transmission of the Buddha's teachings from one teacher to the next without compromising the integrity of the teachings. This transmission takes place through the practice of a lung (see below) transmission.

Lung

Oral transmission of a Buddhist text by a Lama (spiritual teacher). This transmission is carried out by reading out the text in its original to the disciples. The Lama who carries out the oral transmission must be a lineage holder (see above) and authorised to do so by the head of the lineage.

Mahayana

The 'great vehicle', referring to those practicing on the bodhisattva path to free all sentient beings from suffering

[1] These terms have been adapted and simplified based on Thaye (2001) for a non-specialist audience.

Sangha

1) Teachers and beings who have achieved realisation, 2) The community of fellow Buddhist practitioners

Shastra

A text that provides a commentary on primary literature around one particular Buddhist philosophic theme or topic

Shedra

A 'centre of teaching', where Buddhist scriptures and texts are taught and studied as part of a monastic curriculum

Shedra in the west

A 'centre of teaching' to study Buddhist scripture and texts in the West by laypeople

Sutras

A collection of texts of teachings spoken by the Buddha. The second of the three 'baskets'

Tibetan Buddhism

Buddhism as taught and practiced in Tibet from around the 7th century CE based on Mahayana and Vajrayana practices

Vajrayana

The 'diamond' vehicle of Tibetan Buddhism, otherwise known as the 'secret' vehicle, which uses mantras, recitation and visualisation to reach enlightenment quickly.

GLOSSARY OF CORPUS LINGUISTICS TERMS

Collocation

Words that co-occur more frequently than would be expected by chance. Calculated in the present study using the two-level statistical measure of mutual information (MI) and log likelihood available in AntConc (Anthony, 2019). MI identifies the collocational strength. It determines collocation by measuring "the frequency with which collocates occur together as opposed to their independent occurrence" and thus MI "will give a high collocation score to relatively low-frequency word pairs" (Baker, Hardie, & McEnery, 2006, p. 38). Log likelihood (LL) provides an additional measure to test the statistical significance of the collocates by comparing the difference in observed frequencies and expected frequencies within the corpus, at a confidence level of $p<0.05$.

Concordance

Also referred to as KWIC – Key Word in Context. A concordance provides contextual information of a node. Concordance lines have been extracted from the corpus in this study using AntConc (Anthony, 2004)

Corpus

A corpus is "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (Sinclair, 2005, section 10) using computer software.

Keyword

Keywords are words that are unusually highly frequent (positive) or unusually highly infrequent (negative) in a corpus when said corpus is compared to another corpus (reference corpus).

Keykeyword

Keykeywords are words that are keywords (see above) in most of the texts contained within a corpus. In the present study, keykeywords were calculated to identify only those words as keykeywords that are key in all corpus texts. Keykeywords were calculated manually in this study (Scott, 2010)

Lexical closure / lexical saturation

Lexical closure, also referred to as lexical saturation, will be used in this study to indicate the lexical representativeness of the corpus, an approach proposed by McEnery and Wilson (2001). Lexical closure is measured by calculating and analysing the lexical growth in the number of types in a corpus as segments of 1,000 tokens are gradually added to the corpus. Closure is indicated where the number of types "tapers off" gradually as more segments are added (Temnikova, Baumgartner et al., 2014a) Thus, a corpus can be shown to achieve saturation, or closure, lexically speaking.

Lexical repetition

Lexical repetition is measured by calculating the type-token ratio of a corpus: the number of types (unique words) is divided by the number of tokens (total number of words) of a corpus, with the result expressed as a percentage (Baker et al., 2006, p. 162). Such a measure can indicate the degree of repetition within a corpus, and thus "a high type/token ratio suggests that a text is lexically diverse, whereas a low type/token ratio suggests that there is a lot of repetition of lexical items" within a corpus. This also causes larger corpora to show a tendency towards a lower type/token ratio, "due to the repetitive nature of function words" (Baker et al., 2006, p. 162).

Multi-dimensional analysis

Multi-dimensional (MD) analysis is a framework for the analysis of genres of speech and writing, developed by Biber (1988). The framework identifies linguistic features associated to five dimensions: involved vs. informational, narrative vs. non-narrative, elaborated vs. situation-dependent reference, overt expression of argumentation and impersonal vs. non-impersonal style (Biber, 1988). The dimensions of the MD analysis have provided a reference point in this present study to help position the genre of shastra alongside other genres of written or spoken discourse.

N-gram

An n-gram is a sequence, or cluster, of *n* words in a corpus. N-grams have been identified in the present study using AntConc (Anthony, 2004)

POS-tag

Part-of-speech annotation of a corpus.

Reference corpus

Frequency based corpus analysis requires a dataset against which the corpus data can be compared. This dataset is another corpus, a reference corpus, which is normally a large general corpus (Baker et al., 2006). In the present study the British National Corpus (BNC) and also the British Academic Written English (BAWE) corpus. Comparing a corpus against a reference corpus will allow for the generation of a keyword (see above) list. The information within a keyword list will depend on the reference corpus that has been selected.

CHAPTER 1: INTRODUCTION

This introductory section will outline the aims of the study, which investigates Tibetan Buddhist English, and will provide a rationale behind such aims by means of a brief background as well as a summary of the pertinent existing literature. It will further illustrate the approach that has been taken to investigate Buddhist English and indicate the contributions that such a research project will make to the wider discipline of Applied Linguistics and, more specifically, English for Specific Purposes. It will further indicate the overall structure and content of this thesis.

1.1 Aim of the study

The aim of this study is to identify the most frequent lexical features of the Tibetan Buddhist genre, Shastra, through the lens of register analysis, and identify how such features compare against general written and academic written registers. Analysis of such data will focus on the functions and implications of such findings within the register's situational context.

1.2 Background of the study

The initial motivation for the present study is based on the personal experience of the researcher with the target culture, and the difficulties encountered. When first engaging with Buddhist communities through meditation groups and through reading Buddhist literature, I observed a highly frequent use of Sanskrit loanwords as well as non-standard use of English, to the extent of experiencing great difficulty in following arguments or conversations. Long-standing community members, it appeared, use this "hybrid" language with specialised terminology naturally without giving it much consideration, but for an outsider like myself, this language use became incomprehensible at times. This experience triggered the motivation for a systematic investigation into the language as it is used in this specific context.

This motivation was further driven by an encounter with monks accompanying Buddhist leaders, such as the entourage of His Holinesses, Karmapa, the Dalai Lama and Sakya Trizin on their visits to the UK. When attempting conversations with the monks, it became quickly apparent that their English language skills were very much constrained to a general usage, and they were unable to communicate about Buddhist concepts. I investigated this further and noticed recent developments in the East (predominantly Nepal and India), where monasteries are increasingly internationalising their reach to Western audiences through the introduction of English to the monastic curriculum. The focus of such efforts has, however, been predominantly on LGP (Language for General Purposes) rather than LSP (Language for Specific Purposes, i.e. Dharma or Buddhist English) that would enable them to communicate with outsiders about their belief. Particularly given the increased interest in Buddhism in the West, and the funding reliance of monasteries in the East on Western benefactors, this is a development that is likely to become essential.

1.2.1 Changing the focus of the thesis

At its outset, given the above considerations, the project intended a practical application of the theoretical insights into the language as it is used within the specific context of Buddhism by means of field work in a nunnery in Kathmandu, Nepal, in order to develop and test Buddhist English learning materials that could be utilised by novice/untrained native language teachers who frequently visit monasteries and volunteer by offering their native English ability to provide language classes to monks. This focal point of the thesis was forced to change as fieldwork was expected to take place in Nepal shortly after the Earthquake in 2015, rendering such investigation inappropriate given the hardship that was encountered by monasteries and the wider communities in Nepal at that time. As the project progressed, it also become apparent that such an investigation would exceed the scope and feasibility of this PhD thesis. It is for this reason that the study presented as part of this thesis is pedagogical in its aims with an indication of how such insights could be applied to materials development in the context of language teaching presented in Chapter 8: Implications.

1.3 Building on existing research

As has been stated above, this study has been driven by the personal experience of the researcher, which, to a large degree, was characterised by frustration with the incomprehensible language use, where members of the target culture would make frequent use of Sanskrit loanwords, a practice that was also mirrored in the scholarship encountered. A first account of the language used within this context coined the term “Buddhist Hybrid English”, highlighting the limitations of texts that have entered the English language by means of translation from Sanskrit or Classical Tibetan, resulting in a “dialect comprehensible only to the initiate”, further characterising the continuation of translations as a contribution to the “bastardization of English”, based on the use of “technical terminology”, largely left unexplained, or used inconsistently (Griffiths, 1981, p. 17).

Further research into Tibetan Buddhist English has proven limited in number and based predominantly on introspection, the analysis of single texts or texts produced by a single author. The focus of such scholarship has been exclusively within the discipline of Buddhology or translation studies (whereby much of the body of work within Buddhology is based on translational research). The growing body around the use of religious language and theolinguistics[2] has widely ignored Buddhism.

The present study aims to fill this gap in the literature by expanding on the work that has been done on Buddhist English by approaching the subject through a register lens and by means of a data-driven corpus approach. The study aims to make a contribution by applying and expanding on Douglas Biber’s work on register analysis (1988, 2007), by utilising his framework of situational, linguistic and functional analysis (Biber & Conrad, 2013) and thus providing insights into linguistic features (or register features) of Buddhist English, as represented in the written subgenre of *shastra* of Tibetan Buddhism. This will allow a positioning of the Tibetan Buddhist subregister of shastra among general written and academic written registers more specifically.[3] Such positioning will make a valuable contribution to existing research where religious texts have not previously been positioned alongside academic written registers.

[2] Arguably, Buddhism does not fall under the category of “theism” as it rejects the concept of a creator god. It has been argued that it can be practised as a religion but based on the teachings by the Buddha, it has been argued, it may perhaps be more appropriately classified as a philosophy (Siderits, 2007) or even cognitive psychology (Segall, 2003)

[3] This decision is based on a pilot study that was carried out in the early stages of the PhD research, that would enable the researcher to gain some early insights into the language use within the genre of *shastra*. This was conducted by means of using corpus linguistics methods on one of the corpus texts included in the present study (full-text): “The Jewel Ornament of Liberation”. This pilot study indicated that the language used (as indicated in the word list generated) appeared similar to the language used in academic writing.

1.4 Method of investigation

The investigation will be carried out by taking a corpus approach to the selection and analysis of data. This approach has been chosen as no previous data-driven, corpus-based research on language use in Tibetan Buddhist shastras has been carried out. As such, this study cannot rely on previous findings to inform the investigation beyond the “Buddhist Hybrid English” hypothesis (Griffiths, 1981), whereby Buddhist language has been described as a “dialect comprehensible only to the initiate, written by and for Buddhologists”.

In order to test the accuracy of such a hypothesis and, more importantly, to explore the typical features of Tibetan Buddhist English, the present study applies a specialised corpus approach to the study of such language, which involved the compilation of the Mikyo Dorje Shedra of Tibetan Buddhism (MDSTB) corpus. The corpus sample was based on external criteria, namely the discourse community that uses such language and is thus comprised of full texts of the shastras that are studied as part of the Mikyo Dorje Shedra curriculum within the Karma Kagyu School of Tibetan Buddhism. As such, the corpus is small in size (281,290 tokens) and imbalanced in terms of text length within the corpus. As a first data-driven linguistic analysis of the language of Tibetan Buddhism through corpus linguistics tools, the full-text inclusion was deemed crucial so as not to exclude pervasive features that may only occur within specific parts of the text.

Data analysis will be based on Biber and Conrad’s analytic framework, and considers: situational analysis, linguistic analysis and functional analysis (Biber & Conrad, 2013). Pervasive features will be identified by means of word frequency lists, keywordlists and keykeywordlist, and further investigations will consider data through frequency of distribution (dispersion), collocation, concordances and full text view.

Findings of the linguistic analyses will be aligned with Biber’s Dimension 1, as part of his multi-dimensional (MD) analysis[4] (Biber, 1988) or the comprehensive work of the Longman Grammar of Spoken and Written English (LGSWE) (Biber, Johansson,

[4] For an explanation of Dimension 1 and MD analysis, see chapter *3.4.2 Multi-Dimensional (MD) analysis*

Leech, Conrad, & Fineagan, 1999). Such alignment allowed the positioning of the written subregister shastra alongside other written registers. Particularly the third aspect of this analytic framework, the functional analysis, will be able to draw on vital contextual information about the use of texts in their Buddhist context to help understand the significance of linguistic features within the corpus.

1.5 Organisation of this thesis and contributions

The present thesis will be organised in the following way. Given the specialised domain of Tibetan Buddhism, and the assumption that key concepts and an understanding of Tibetan Buddhist traditions is widely unknown to a general audience, a brief introduction into Tibetan Buddhism and Buddhist education will be provided in chapter 2. This will also indicate the broadness (or narrowness) of the focus of this thesis and position the register under investigation within its wider discipline.

Chapter 3 will indicate the linguistic investigations into the English language use within the context of Tibetan Buddhism that have been carried out, at large drawing on the work of Griffiths (1981) and his claims of the “incomprehensibility” of language in Buddhist English written registers. An investigation of the literature around corpus linguistics in general and specialised corpus approaches to conduct data-driven empirical research will be presented, and it will be indicated that there is a lack of such strategic data-driven investigations into the language use within the Buddhist context. Having established this gap in the literature, the chapter will culminate in identifying suitable analytical frameworks for the analysis of Tibetan Buddhist English Shastras, and thus illustrate the affordances that the multi-dimensional (MD) analysis framework (Biber, 1988) and the framework of situational, linguistic and functional analysis (Biber & Conrad, 2013) for the study of registers can contribute to the present project.

Significant contributions to the wider discipline will be made methodologically through chapter 4, which justifies the creation of and thought processes underpinning the small specialised corpus. The data provided through this corpus will enable the researcher to apply existing work within small specialised corpora to the context of Buddhist English, more specifically to the academic written subgenre of shastras. Such analysis will enable the formulation of a description, classification and analysis of the interplay between linguistic features and situational context of shastras. In this way, the research questions formulated at the outset of chapter 4 will be addressed: (1) What are pervasive linguistic features of the written subregister shastra in Tibetan Buddhist English?; (2) In relation to question 1, what are the characteristics of such linguistic features?; (3) What is the link between such linguistic features and their situational context of Tibetan Buddhist shastras?; and (4) How do the linguistic features of Tibetan Buddhist shastras compare to other written registers?.

Chapter 5 is the first analysis chapter and will provide an insight into the “aboutness” of the Tibetan Buddhist shastras through analysis of its lexical closure properties, lexical repetition and word frequency lists. Calculation of lexical closure will be used as a measure of lexical representativeness. Despite its relatively small size, the Mikyo Dorje Shedra of Tibetan Buddhism corpus achieved near closure: as new texts were added to the corpus, the number of types remained (nearly) unchanged. The closure properties of the corpus will be indicated in comparison to general written registers (BNC) and academic subregisters (CRAFT corpus). It will thus be argued that the corpus is lexically representative of the language it represents. In this way, the present study will contribute to the wider methodological discussion around representativeness in corpora, particularly for those small in size aiming to represent a specialised subregister.

The second part of chapter 5 will measure lexical repetition of the MDSTB corpus by means of type-token ratio (TTR), and indicate how its unusually low TTR is accounted for by the closure properties of the corpus on the one hand, and a frequent repetition of headings and subheadings within the corpus on the other. Contextualising such findings with the Buddhist practice of memorisation and debate, the argument will be made that the low TTR is a direct result of the way language in the texts is used as mnemonic devices to aid the memorisation of such texts. Such analysis, it will be argued, is indicative of the limitations of quantitative-only analyses, and the case will be made for the contribution that the application of Biber and Conrad’s (2013) framework for the analysis of registers can provide, thus filling the gap between contextual information and linguistic analysis.

The final section of chapter 5 will investigate the language of Tibetan Buddhism through word frequency lists (wordlist, lemmatised lexical wordlist and keykeywordlist), and utilise such findings as the starting points for further investigations in subsequent chapters.

For example, the investigation into the use of *one* in chapter 6 will be based on findings from the wordlist, where it will be highlighted as an unusually highly frequent token in the MDSTB corpus compared to the British National Corpus (BNC) and British Academic Written English (BAWE) corpora. Analysis of *one* will be carried out at word level to indicate frequency data, as well as at sentence level (and beyond) by means of concordancing. Findings will be aligned alongside other academic written registers by means of comparison with the Longman Grammar of Spoken and Written English academic subcorpus.

The use of *one* as a substitute pronoun, a generic pronoun and its use as part of proper noun (name) will form the basis of this investigation. Such use, it will be argued, is common in the specific context of Buddhism, and it will be illustrated that the use of *one* in this way is indicative of the translational practice of using literal translations of loanwords in Buddhist English, thus posing the challenge of comprehension of Buddhist shastras for an audience outside the target culture.

The use of *one* in its use as a substitute pronoun commonly functions as an anaphoric countable reference (e.g *this one*) and is as such most frequently used in oral conversation and less frequently in written registers. It will be illustrated that the unusually high frequency of *one* in this context not only functions as a cohesive device in general, but more specifically utilises the “countable referencing properties” of this device to cross-reference to concepts that are comprised of components which are numbered. As such, the argument will be made that its use refers to the Buddhist practice of memorisation as presented in chapter 5.

The use of *one* as a generic pronoun will consider its use as part of different syntactic functions (i.e. sentence object or subject), and its use as the subject within conditional subordination structures. It will be shown that the use of *one* is centred around the topic of “cause and effect” and it will be argued that its use as the sentence subject provides agency to *one* and as such implicitly communicates empowerment to the reader. Further, it will be argued that the use of *one* will indicate objectivity and thus universal applicability of the content communicated and overcome the dualism that can easily be inferred from substituting *one* with personal pronouns. As such, it will be argued that the use of *one* within the corpus becomes a vehicle to reflect the Buddhist thought which the written subregister shastra aims to communicate. The main contribution based on this chapter will be empirical in that the investigation into the use of *one* will enable the researcher to position the shastra alongside other written registers. Simultaneously, its use of *one* as an anaphoric referencing device*,* here in the context of a formal written register, will challenge the theoretical assumption that a high frequency of the substitute pronoun *one* is indicative of spoken informal registers (cf. Biber, 1988).

Chapter 7, the final analysis chapter, will investigate the use of Sanskrit loanwords in the Mikyo Dorje Shedra of Tibetan Buddhism corpus. The first part of the chapter will investigate spelling variation within the corpus and link such findings to the transfer of loanwords and proper nouns (names) from Sanskrit into English. The second part of the chapter will provide a case study of the Sanskrit loanword *bodhicitta*, a technical term to denote the Buddhist concept of the aspiration to reach enlightenment for the benefit of all beings. Analysis of *bodhicitta* will consider the use of synonyms, near-synonyms and variants, and subsequently, based on Sinclair (2004), consider semantic prosody, collocation and semantic preference of *bodhicitta.*

Collocation analysis will indicate a preference of bodhicitta to collocate with CULTIVATE and compare such findings to collocates of CULTIVATE within the BNC written corpus, indicating similarities in such use in terms of the meaning of “cultivating an attitude of mind”, a concept that very much resembles the use of bodhicitta in its use to demarcate an intention or aspiration.

Concordance lines will be analysed to help classify and provide a broad understanding of the concept of bodhicitta through its representation within the corpus. As such, corpus tools will not only provide a useful tool for the analysis of data but also for the unravelling of meaning. Despite the insights that can be gained into the concepts of Buddhism by means of corpus tools, it will be argued that the language use at large is challenging to comprehend, based on the high frequency of loanwords. The main barrier to comprehension, it will further be argued, is the use of different terminology to denote the same or similar concepts, as well as specific loanwords denoting different concepts. As such, these empirical findings will build on the argument of the incomprehensibility of Buddhist language put forward by Griffiths (1981), furthering his argument methodologically through the employment of corpus methods, and thus providing a systematic, data-driven account of his hypothesis that was largely based on single texts in translation.

By linking linguistic analysis with situational analysis (Biber & Conrad, 2013), the final section of this chapter will, however, deviate from this argument. The Buddhist practice of textual study as part of “Shedras in the West” (where texts are taught in lecture-style events, with commentary and contextual information provided, followed up by small study groups), indicates a mitigation of the miscommunication of Buddhist concepts by means of oral communication, that would persist if texts were studied as a “stand-alone” by a non-specialist individual, without an understanding of Sanskrit or Buddhist philosophy more widely.

Chapter 8 will consolidate implications that arise from the present thesis. In the first section, the methodological implications will be highlighted through reflection of the process of compiling and analysing a small corpus. Templates for other researchers wishing to apply a similar approache will be provided and cover: compiling a small corpus, measuring representativeness through lexical closure and corpus analysis through situational, linguistic and functional analysis (Based on Biber and Conrad, 1993). The second part of this chapter will illustrate how corpus findings can be pedagogically applied to the creation of learning materials, based on the principles of data-driven learning.

The thesis will conclude in chapter 9, where the main arguments will be summarised, with a particular focus on the empirical, methodological and theoretical contributions made as part of this thesis. Limitations to the findings of this thesis will be highlighted and recommendations will be made on the basis of such for further research.

CHAPTER 2: TIBETAN BUDDHISM AND BUDDHIST EDUCATION

2.1 Introduction to the chapter

As the present study is a register analysis into the written Buddhist genre[5] of shastra, as studied within the Kagyu school of Tibetan Buddhism, it is essential to provide the reader with a broad overview of what constitutes Tibetan Buddhism and its four schools.[6] This matter is, however, a complex one, historically and philosophically, and given the nature of this investigation, being an applied linguistics PhD thesis, a detailed account that would do the subject justice is beyond the scope of this study.[7] The following broad, somewhat crude, overview serves the sole purpose of merely providing the reader with a general understanding of the context of the topic to enable a contextualisation of the Buddhist tradition under investigation.

This section will provide a broad overview of Tibetan Buddhism, Tibetan Buddhist education in the East, its growth in the West and the general and linguistic implications of this process of globalisation. Drawing on this, the final segment within this section will highlight the necessity of an investigation into the linguistic properties of Tibetan Buddhism and briefly touch upon its applications in the context of Buddhist Education as well as translation work in the West.

2.2 Globalisation of Buddhism

Within the last forty years, Buddhism has spread globally, leading to increasing numbers of practitioners world-wide (Kölling, 2011). Due to the occupation of Tibet in the 1950s, monks and lamas have founded new monasteries in exile in Nepal and India. This development has led to the existence of transnational religious communities within Tibetan Buddhism. Reasons provided for this increasing interest

[5] Where the term *genre* is used in this thesis, it denotes a specific Buddhist text type and is not indicative of the approach or linguistic analysis conducted in this thesis

[6] The present study adopts the perspective of differentiating between different *schools* of Tibetan Buddhism: Nyingma, Gelug, Sakya and Kagyu. Another way of differentiation would consider *principal streams of spiritual practice*, which distinguishes between eight or even nine so-called “chariots of practice” (Thaye, 2013).

[7] A comprehensive account of Tibetan Buddhist traditions has been provided elsewhere (Stott, 1980).

has been ascribed to increasing immigration from Buddhist countries which gave rise to Buddhist lamas receiving increased media attention (Wuthnow & Cadge, 2004, p. 364). Other research (Kölling, 2011) attributes this increase to Tibetan Buddhist monasteries pleading for support from the West to enable the re-building of their monasteries in exile. Tibetan lamas have increasingly travelled to the West, and since the 1970s there has been an increase of *dharma* centres in the West. In the present day, the number of Buddhist laypeople has come to exceed the number of ordained monks, leading to lamas spending significant time abroad on so-called "teaching tours" to instruct practitioners globally. This expansion of Tibetan Buddhism has led to a mutual influence between East and West within Tibetan Buddhism (Kölling, 2011).

2.3 Buddhist terms

2.3.1 Tibetan Buddhism

Thaye (2001) provides an excellent, yet concise overview of what constitutes Tibetan Buddhism. Broadly speaking, the term 'Tibetan Buddhism' is a Western concept, coined to demarcate Buddhist practice in "Tibet and in the surrounding areas such as Mongolia, Bhutan, Ladakh, Sikkim, parts of Nepal and even parts of Siberia" (Thaye, 2001, p. 2). Further, one of its key characteristics is the "notion of *lineage*". This aspect of Tibetan Buddhism will be of particular importance in later sections of this thesis that outline the rituals surrounding textual studies. In essence, a *lineage* entails the "unbroken transmission" of the teachings given by the Buddha from teacher to disciple, where a teacher, also called *lama*, is empowered to transmit texts only upon mastery of such.

At the core of the fundamental teachings of the Buddha lies the path that leads to the liberation from suffering, also called "awakening" or "enlightenment". The teachings can be divided into two strands through focus on motivation. The *hinayana*, also called "ordinary teachings" or "Lesser Vehicle" encompass those teachings aimed at achieving liberation from suffering for oneself. The *mahayana*, also called "extraordinary teachings" or "Great Vehicle", are the teachings aimed at achieving liberation from suffering for all beings, not just for oneself as in the *hinayana*. The practitioner on the path of the *mahayana*, driven by compassion for others, is called a *bodhisattva*. The *mahayana* itself includes “a further set of teachings”, “known as the *tantras*” which led to the “doctrine known as *vajrayana*” (Thaye, 2001, p. 7).

Within each of these strands, the teachings of the *hinayana* and *mahayana* comprise different textual collections. The teachings of the *hinayana* are grouped into three so-called “baskets”: the *abidharma* (philosophy), *vinaya* (monastic conduct), and the *sutras* (discourses[8] of the Buddha). The teachings of the *mahayana* were spread predominantly by the Buddhist masters Nagarjuna and Asanga who, amongst others, “contributed to the systematization of the […] teachings” and “expanded, defended and codified” those teachings (Thaye, 2001), leading to the establishment of the philosophical schools of the *mahayana*: the *madhyamaka*, also called “Middle Way”, and the *chittamatra*, also called “Mind Only” school.

In addition to the two philosophical schools, one can also differentiate two “divisions” of the *mahayana* according to the practice outlined within the texts: the *ordinary* and *extraordinary mahayana*. The practice of the *ordinary mahayana* is a “graduated form of practice” and also called the *paramitayana*, the ‘Vehicle of Perfections’. The *extraordinary mahayana*, as the name suggests, is a “skilful, rapid form of practice” also called the *mantrayana*, the ‘Way of Mantras’, and even more commonly referred to as *vajrayana*” (Thaye, 2001, p. 9). The *vajrayana* in itself contains four sets of *tantras* with each set of tantras being grouped according to the subject matter contained therein: the *kriya tantras* (ritual practice), the *charya tantras* (ritual practice with stronger emphasis on meditation), the *yoga tantras* (focus on meditation alone) and the *anuttara tantras* (subtle means of acquiring realisation not contained in other tantra sets) (Thaye, 2001).

The complexity of the teachings of the Buddha is indicative of the multitude of paths that enable practitioners to free themselves from suffering, in other words, to reach enlightenment. The provision of different paths enables the *dharma*, the teachings of the Buddha, to cater for different practitioners’ characteristics and personality types. A visual overview of the above classifications has been provided in *Figure 1: Classifications of the Teachings of Tibetan Buddhism* on the next page.

[8] Discourses in the context of Buddhism refers to the teachings of the Buddha

Having broadly categorised Tibetan Buddhism in terms of motivation, philosophical schools and different paths in the previous section, the following paragraphs will position the *Kagyu School of Tibetan Buddhism*, the school under investigation within this study. This school has been chosen as a focal point as it is the most widely searched for traditions in Europe in the present day and overall world-wide (Google Trends, 2020), and the researcher has already established networks within the Kagyu School that can inform the data collection for the present study.

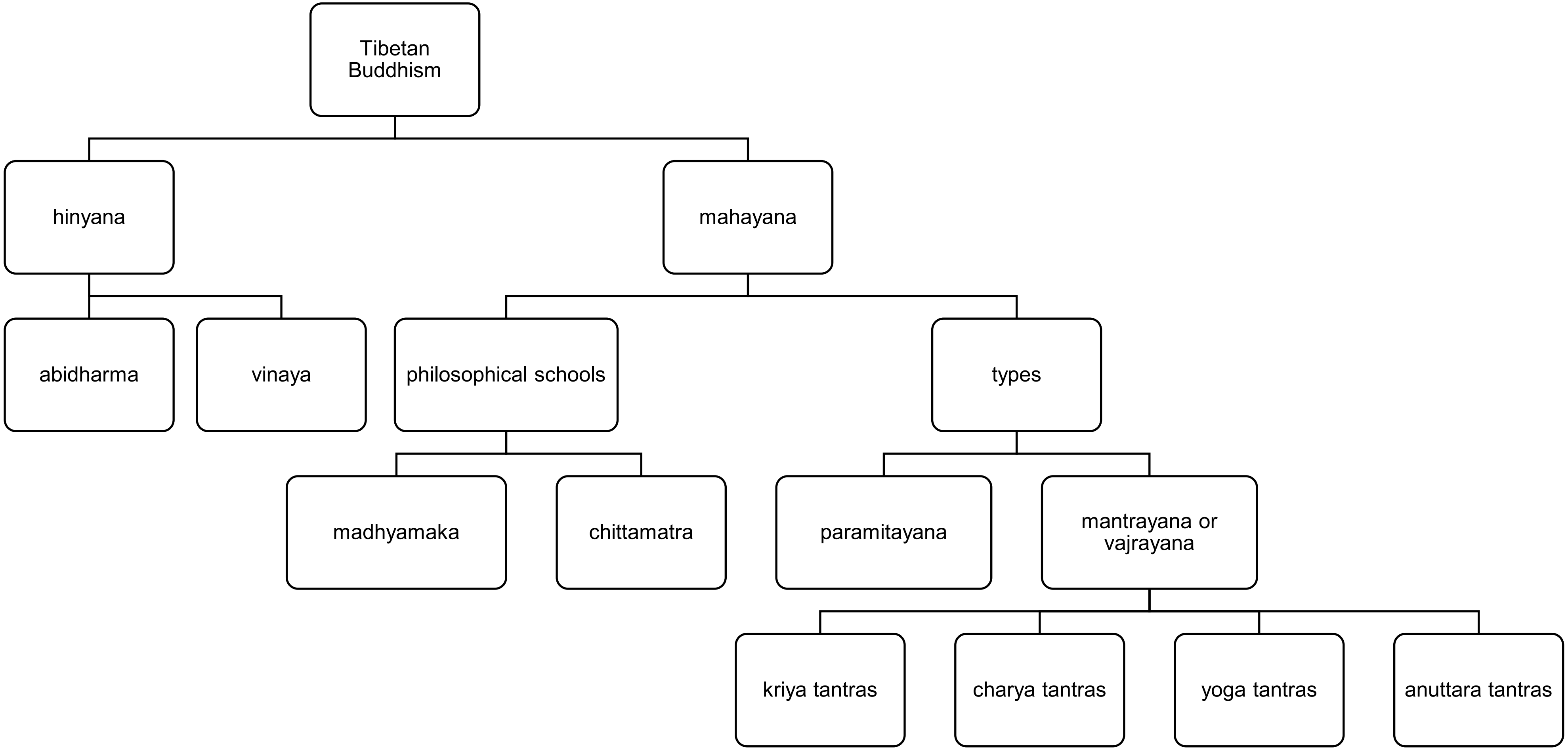


Figure 1: Classification of the Teachings of Tibetan Buddhism

2.3.2 Overview of the schools of Tibetan Buddhism

In order to understand why there are different *schools* of Tibetan Buddhism, one has to understand, in very general and broad terms, how the schools developed historically. As described in an earlier section of this chapter, one of the key characteristics of Tibetan Buddhism is the "notion of *lineage*" (Thaye, 2001, p. 72), the unbroken transmission of the teachings given by the Buddha from teacher to disciple (see 1.1.2.1). Additionally, the variety and multitude of teachings unsurprisingly lead to certain texts being studied, practiced and transmitted predominantly or even exclusively by specific groups of practitioners and teachers. This, in turn, as can be seen from the following citation, led to the development of different schools of Tibetan Buddhism.

> Buddha had given so many different practices to lead beings to enlightenment that clusters of gurus and disciples were able to specialize in one particular set of teachings, for which perhaps they alone held the transmission. Thus the schools represent the culmination of a long process of development carried on by such groups each of whom transmitted a series of instructions from the Indian inheritance (Thaye, 2001, p. 72).

As a result of these "groups" or "clusters", four different schools of practice arose in Tibet: the *Nyingma*, *Sakya, Kagyu* and *Gelug* schools of Tibetan Buddhism. This section will merely touch upon their origination to enable the reader to grasp the core differences between the schools.[9] An overview of the Schools of Tibetan Buddhism has been provided in *Figure 2: Schools of Tibetan Buddhism*.

2.3.2.1 Nyingma

[9] As each of these schools have existed for over a millennium, significant changes and influences have taken place which shall not be outlined in this section as they are deemed irrelevant to the present study.

The *Nyingma* school is the oldest of the four schools and originated in the late eighth century, and emphasis is given to the tantric teachings of Padmasambhava, a tantric master also called the “Precious Guru”. The tantras studied and practiced are also referred to as the “ancient” tantras. The three schools of Tibetan Buddhism that originated subsequently are referred to as the “Schools of the New Tantras” (Thaye, 2001, p. 80).

2.3.2.2 Sakya and Kagyu

Both, the *Sakya* and the *Kagyu* school originated in the eleventh century and have their roots in the *Nyingma* school. The *Sakya* school was formed by practitioners of the ancient tantras in order to distance themselves from practitioners of the ancient transmission who they considered to have become “compromised by the dubious behavior of some of their adherents” (Thaye, 2001, p. 81). Subsequently, they received teachings on the *Hevajra tantra*. Following the influential work by Konchog Gyalpa, the Sakya tradition transformed from a small group of practitioners to “one of the greatest schools that the Buddhist world has ever seen” (Thaye, 2001, p. 81). The *Kagyu* school was founded by Marpa (also called ‘the translator’), Milarepa (an ascetic yogin and poet) and Gampopa (a monk and scholar). Broadly speaking, the *Sakya* school places great emphasis on “scholarship and tantric ritual”, the Kagyu school on “meditation and yogic practice” (Thaye, 2001, p. 85). Within the *Kagyu* school of Tibetan Buddhism, several sub-sects developed, such as the *Karma Kagyu* school.

2.3.2.3 Gelug

The *Gelug* school, with its origins in the late fourteenth, early fifteenth century, is the most recent of the four schools and, like the *Sakya* and *Kagyu* schools of Tibetan Buddhism, belongs to the new tantra schools. It was founded by Tsongkhapa from Amdo who synthesised the philosophical and meditative teachings of the other schools with a strong focus on the “graduated path” (Thaye, 2001, p. 92). Additionally, the educational curriculum of the Gelug gave a strong emphasis on the study of logic and epistemology.

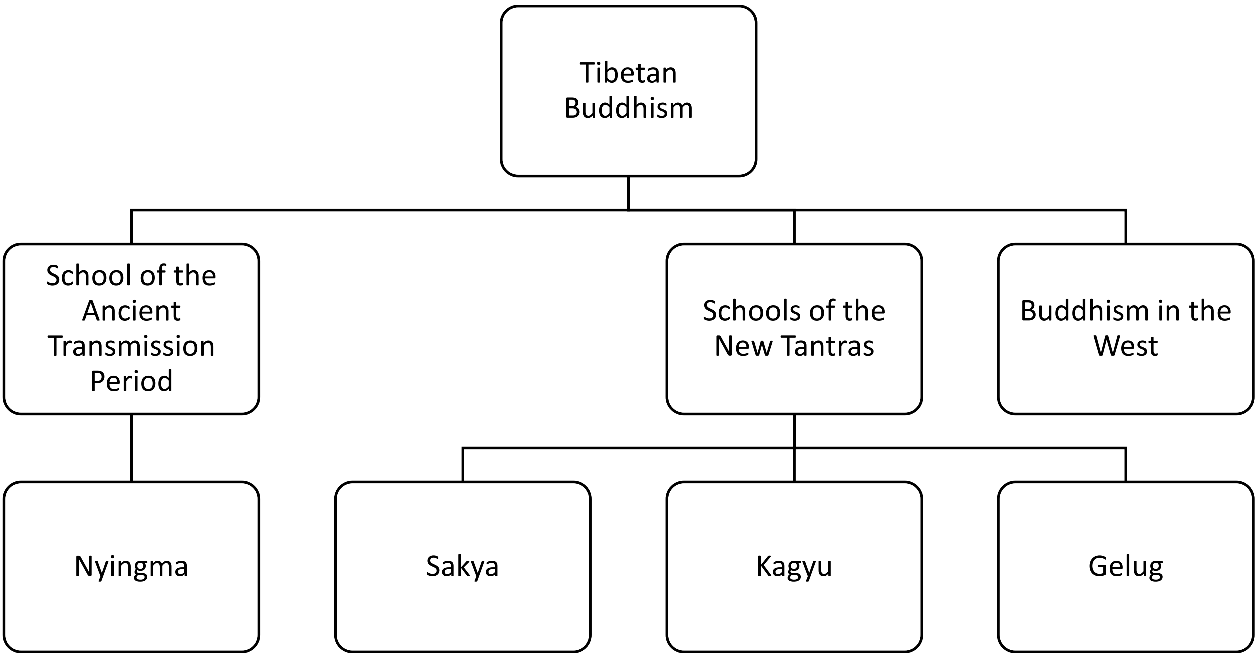

Tibetan Buddhism
School of the Ancient Transmission Period
Schools of the New Tantras
Buddhism in the West
Nyingma
Sakya
Kagyu
Gelug

Figure 2: Schools of Tibetan Buddhism

To a Western non-Buddhist audience, the Gelug school of Tibetan Buddhism is perhaps the most well-known one, as, in the seventeenth century, Ngawang Losang Gyamtso, the fifth Dalai Lama, became ruler of Tibet, making the Gelug school of Tibetan Buddhism essentially the state church of Tibet (Thaye, 2001, p. 93). In his current rebirth, Jetsun Jamphel Ngawang Losang Yeshe Tenzin Gyatso, the fourteenth Dalai Lama, is arguably the person most widely associated with Buddhism and Tibetan Buddhism in the West.

2.3.3 Summary

This section has provided a very broad classification of Tibetan Buddhism. These classifications are by no means exhaustive but allow the researcher to indicate the location of the texts under investigation in the present study. The texts form the curriculum as studied in monasteries of the Kagyu school of Tibetan Buddhism. Further information about the textual selection will be provided in chapter 3: methodology.

To briefly summarise this section, Tibetan Buddhism can be geographically and historically linked to Tibet as well as Mongolia, Bhutan, Ladakh, Sikkim, bordering regions of Nepal and parts of Siberia. Tibetan Buddhism is part of the v*ajrayana* tradition of Buddhism, which in itself is a subsection of the *mahayana* tradition. The motivation underlying the latter tradition is the achievement of liberation from suffering for all beings rather than oneself with the *vajrayana* subsection placing main emphasis on the teachings given in the tantras. The four schools of Tibetan Buddhism, given their origination, have strong overlaps in their practice and philosophies but differ in terms of emphases on texts and thereby practices. The Kagyu school, as one of the four schools of Tibetan Buddhism, has a strong focus on meditative practice which is reflected in the teachings that underpin this tradition, and will be the sole focus of this study.

2.4 Buddhist Education

Having previously outlined what constitutes Tibetan Buddhism and its philosophical underpinnings, the present section is concerned with the culture of studying such philosophy.

2.4.1 Shedra curriculum of the Kagyu School of Tibetan Buddhism

Places for the study of Tibetan Buddhism are so-called s*hedras*, broadly defined as “religious centres” (Phuntsho, 2000, p. 98) or “monastic universities or colleges” (Kölling, 2011, p. 18) or, more specifically, “a college of studies attached to a major monastery” (Dechen, 2008). Topically, such Buddhist educational institutions would provide a broad overview of Buddhist philosophies, appropriate for the Kagyu School of Tibetan Buddhism. As in the transmission of lineages, it is of importance that such instruction is provided by a recognised master of such philosophies:

> There [in *shedras*] they [the monks] would be able to pursue studies under the direction of highly qualified masters on the topics of Madhyamaka (The Philosophy of the Middle Way), Paramita (The Path of the Bodhisattva), Pramana (Logic and Epistemology), Vinaya (Monastic Discipline), Abhidharma (Higher Teachings) and Tantra. (Dechen, 2008, para 4)

The purpose of such education was the provision of an in-depth understanding of Buddhist philosophy and thereby a “superb foundation for meditation”. Students who successfully completed their *shedra* studies may subsequently become teachers of such philosophical systems themselves (Dechen, 2008, para 5). *Shedras* thereby not only play a vital part in the dissemination of Buddhist philosophy but also the survival thereof in line with the principles of uninterrupted transmission from lama to disciple.

2.4.2 Internationalisation of monasteries in the East and Shedras in the West

An increasing interest in Buddhism in the West also led to the introduction of *Shedras* in the West, as well as Buddhist Universities or colleges in the East for Western disciples. These aim at providing understanding of Buddhist philosophies to Western laypeople but also to increasingly forming the foundation for further translational studies.

The textual genre that is studied in shedras and that communicates such philosophies are called *shastras*. Exactly which shastras are studied depends on the school and on the educational institution. All shedras broadly cover the same topics but the translations chosen may vary. Thus, the texts that are studied are based on a canon, and within the textual selection of canonical literature, there is discrepancy. This issue will be followed up in more detail in the methodology chapter of this thesis.

2.5 The emerging written subregister of Buddhist English shastra

English shastras are a very recent addition to the English language, and are based on the translation from Tibetan and Sanskrit. Arguably, based on the observations above, Buddhism is still in the early stages of this transfer of such texts into Western culture, and will yet have a long way ahead in the way such texts are likely to evolve over time:

> Translation occurs most frequently as a result of the transfer of a religion into another culture, but this transfer process has several stages. As a religion develops in a particular area, priorities change. At first, access to the sacred texts is paramount and is often achieved without too much thought about the process of translation. Next comes consolidation of the canon, followed by more

> translations, followed by analysis and justification of translation methodology. Each process can take centuries to evolve and different religions are at different stages in different cultures. (Long, 2013, p. 464)

It is anticipated that many more translations of such canonical literature are to follow, and that the body of Buddhist texts in English will increase, gradually becoming more principled and systematic in the way such translations are produced, and as a result, the language use will become increasingly standardised.

At this moment in time, however, an investigation into the language of Tibetan Buddhism as it is used in the written subgenre of shastra will contribute significantly to the body of work available, in that it will form a first data-driven investigation into such language use and thus provide first insights into the typical features within this register.

CHAPTER 3: LITERATURE REVIEW

3.1 Introduction to the chapter

This chapter will provide an insight into the research into English language use within the context of Tibetan Buddhism, highlighting the limitations in the existing body of literature within this field. A first account of an analysis into the language used within this context coined the term “Buddhist Hybrid English”, highlighting the limitations of texts that have entered the English language by means of translation from Sanskrit or Classical Tibetan, with the result of a “dialect” almost incomprehensible to non-specialist audiences.

Further research into Tibetan Buddhist English is limited in number and based predominantly on introspection, the analysis of single texts or texts produced by a single author. The focus of such scholarship is primarily within the discipline of Buddhology or translation studies (whereby much of the body of work within Buddhology is based on translational research). It will be established that there is a severe limitation in the literature that the present study can draw upon to provide insights into linguistic features (or register features) of Buddhist English, as represented in the written subgenre of *shastra*. The mere linguistic description available is limited to features that cause translational issues, namely the use of loanwords. As such, it is expected that a high frequency of loanwords will be observed as part of the present investigation.

Due to the lack of prior research, the focus of this section of the literature review is to identify other studies that carried out corpus research into written genres and registers, and more specifically into written academic subgenres using specialised corpora. The findings from this synthesis of literature will allow a positioning of the Tibetan Buddhist subregister of shastra among general written and academic written registers more specifically.[10]

[10] This decision is based on a pilot study that was carried out in the early stages of the PhD research, that would enable the researcher to gain some early insights into the language use within the genre of *shastra*. This was conducted by means of using corpus linguistics methods on one of the corpus texts included in the present study (full-text): “The Jewel Ornament of Liberation”. This pilot study indicated that the language used (as indicated in the word list generated) appeared similar to the language used in academic writing.

The final section of this chapter will indicate how findings of corpus research can be practically applied to materials design in the context of language teaching, which will further inform the creation of indicative materials in chapter 9 of this thesis.

3.2 Linguistic investigations into Buddhist English

Linguistic investigations into Buddhism are largely limited to the analysis of Buddhist language use from a translational perspective. As such, literature on Tibetan Buddhism focuses on texts produced in Sanskrit or Classical Tibetan. A first, arguably unfavourable, account of the language used within the context of Buddhist English has been provided by Paul Griffiths (1981, p. 17), who describes Buddhist language as a “dialect comprehensible only to the initiate, written by and for Buddhologists” and further goes on to coin the term “Buddhist Hybrid English” to name said dialect. This terminology is directly derived from the term “Buddhist Hybrid Sanskrit”, which was coined almost thirty years earlier by Franklin Edgerton (1953) to describe the translational issues of introducing Buddhist thought produced originally in Sanskrit into Tibetan. Much of Griffiths’ argument is based upon Edgerton’s work, yet expands upon the results of such translation efforts to the, at the time of his article, very recent productions of English language translations of Buddhist texts:

> Buddhist thought has a strange, and in many respects deplorable, effect upon language; in India it produced that barbaric language we usually call by the equally barbaric name of Buddhist Hybrid Sanskrit, a language in which large numbers of long, repetitive, obscure, and subtle works were composed over a period of more than a thousand years. It forced the Tibetans to invent not only an alphabet but also what was in effect a new language, the most mechanical form of translationese which the world has yet seen. It managed to disturb even the severe balance and precise rhythms of classical Chinese. And it is now in process of wreaking its havoc upon the English language, creating a dialect comprehensible only to the initiate, written by and for Buddhologists, a dialect which has provided the title for this paper: Buddhist Hybrid English (Griffiths, 1981, p. 17)

The “Hybrid” component in “Buddhist Hybrid English” refers to the high frequency of Sanskrit loanwords within the text, and the creation of new terminology, caused by the challenge of translating Buddhist philosophical thought into a language that is based within a different target culture to that where such texts originated, to the degree of incomprehensibility.[11]

Despite the wide recognition Griffiths received for his work, little research has been carried out over the past 40 years since the article was published, to further the insights provided into translated texts in the context of Tibetan Buddhism.

His findings have been predominantly drawn upon within the field of translation studies. Publications that have drawn on the work of Griffiths have sought to overcome the criticisms expressed by him. Yet, even twenty years after his publication, research on the methodology applied in the Buddhist translation practice into English, indicated that many of the issues have remained the same:

> Even today there is no universally agreed method for rendering even some of the most basic and frequently-met technical terms found in Tibetan and Sanskrit into English, or for that matter into any other European language. (Gaffney, 2000)

The reasons for this are predominantly rooted in the lack of expertise available in the West, whereby a translator of such texts needs to be equally well versed in Sanskrit, Classical Tibetan as well as Buddhist philosophy (Gaffney, 2000; Griffiths, 1981).

Additionally, the body of research that has been produced around the use of Buddhist English has been driven by introspection or on the base of a single text (e.g. Blumenthal, 2004) or specific author (e.g. Tillemans & Smith, 1999) as part of the theory or practice of translation, yet research approaches to the language use are lacking a data-driven approach to provide a full account of the language used within English language Buddhist texts, and to enable full accountability of the data within this.

This is how the present study aims to make a significant contribution to the body of research. By taking a corpus approach, this study will consider a specific subregister

[11] Incomprehensibility of Buddhist texts by the author of this study has also been raised as one of the motivating factors for the present study and has been described in the introduction chapter of this thesis.

within Buddhist English, and carry out a first data-driven account of the Buddhist English.

The affordances of a corpus to the investigation of texts in a Tibetan Buddhist context were recently applied in an Arts and Humanities Council funded research project at the School of Oriental and African Studies (SOAS) in the creation of a POS-tagged corpus of Tibetan texts, which will significantly enhance the creation of Tibetan dictionaries (Garrett, Hill, Kilgarriff, Vadlapudi, & Zadoks, 2015). Such efforts indicate important work that is carried out in the context of Tibetan Buddhist language (in first iteration of translation from Sanskrit to Tibetan and its original in Tibetan) and which will undoubtedly influence and help translators address the shortcomings in terms of a standardisation in the practice of translations, and thus significantly influence the language used in the English Buddhist context. As such, the contributions made by the present study must also be seen as a mere snapshot – as are any linguistic analyses as language is ever-evolving – yet, in the context of Buddhism, perhaps more rapidly in the years to come than in others.

3.3 Corpus linguistics and the specialised corpus approach

A corpus is defined as “a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research” (Sinclair, 2005, section 10). Though this definition provides a useful view of the data contained within a corpus, it needs to be furthered to include the analysis of the corpus data, which provides the final insight into the use and function of the corpus data. The advantage of a collection of texts,[12] representing a specific variety of language and available in digital form, is that it can be analysed using computer software, thus enabling the analysis of large textual collections. It can thus be the “starting-point of linguistic description or […] a means of verifying hypotheses about a language” (Crystal, 1991, p. 95). The corpus approach has been chosen for this present study as it

[12] The term “text” in this chapter is used to denote written text as well as transcription of speech but the present study will be limited to written text only.

allows the researcher to analyse the language which is used within the specified context by identifying salient features of said language in use.

The affordance of corpus research, at its outset, is to allow for the identification of frequently occurring items (by means of a wordlist). Such insights can then be investigated further to lead to generalisations to be made on the language use as represented in the corpus (Sinclair, John, 2005, section 10) Such analysis is quantitative in nature, driven by word frequencies as displayed in word lists, or, when compared to another corpus, in the form of keyword[13] or keykeyword lists[14] (positive and negative), i.e. unusually (in)frequent items in comparison to language usage in a reference corpus.

At this point, I would like to highlight Chrystal's word choice of "*starting-point* of linguistic description". Corpus linguistics research that limits itself to mere "number crunching" (Aarts, 2000) has been criticised for decontextualisation of data and lack of critical interrogation of the reasons behind the occurrence of frequencies, limiting themselves to mere description. It is indeed crucial to see the frequency data as a starting point, in particular for the identification of salient features for register analysis. The quantitative dimension will be supplemented through qualitative analysis of the contextual data of single items of the corpus, also called key word in context (KWIC) concordances. Sinclair terms these two complementary dimensions of analysis the vertical (quantitative) and horizontal (qualitative) dimensions:

> The use of a corpus adds quite literally another dimension to language research. If you examine a KWIC concordance, which is the standard format for reporting on recurrence, it is clear that the horizontal dimension is the textual one, which you read for understanding the progress of the text and the meaning it makes as a linear string, while the vertical dimension shows the similarities and differences between one line and the lines round about it. The main "added value" of a corpus is this vertical dimension, which

[13] Keywords are words that are unusually highly frequent (positive) or unusually highly infrequent (negative) in a corpus when said corpus is compared to another corpus.

[14] Keykeywords are words that are key words (see above) in most (or all) of the texts contained within a corpus

> allows a researcher to make generalities from the recurrences.
> (Sinclair, 2005, section 10)

Corpus research is thus helpful not only in automatically displaying reoccurring items based on frequency within a corpus, but also in providing the immediate context in which all occurrences of chosen lexical items occur, thus enabling the researcher to identify patterns of language use within a large collection of texts without the loss of the qualitative aspects of data.

One of the key criticisms of the corpus approach was expressed by Chomsky (1957). His core argument was that any corpus would be limited in that it would not be able to include all features of the target variety as it could not include the variety as a whole. Thereby, some linguistic features, for example low frequency expressions may not be included, or other features may be over/underrepresented in such a corpus which would inevitably lead to its results being skewed. These false representations could be lead back to chance (McEnery & Wilson, 2001). The development of much larger corpora than existed when such criticisms were initially expressed, have led to general corpora that can include significantly larger samples of a target language. Regardless, Chomskyan criticisms still hold up and have been re-uttered by, for example, Charles Fillmore: “I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate” (Fillmore, 1992, p. 35).

In addition to the criticisms around representativeness of a corpus, a practical limitation of the corpus approach is the feasibility of the analysis of highly frequent items within the “horizontal dimension”. The larger a corpus and the more frequent an item occurs, the more time is required for the analysis of said items in a qualitative manner. It is for this reason perhaps, that studies based on large general corpora appear to favour the quantitative over the qualitative dimension, and have faced criticism in their tendency to decontextualise naturally occurring language from its original context into mere word frequencies, thus separating the communicative context in which language occurs from the data itself (Widdowson, 1998). Widdowson’s argument is further supported by Flowerdew (2004):

> In order to fully and accurately interpret the corpus data it is necessary to be cognizant of the role that the context of situation and context of culture play in shaping the discourse under investigation (p.16).

Analyses of general corpora, though they make valid and important contributions to understanding both the use and structure of language, are less useful to provide insights into the use of language within specific contexts, such as the language use within specific academic or professional contexts. Such varieties of language have increasingly been studied by means of specialised corpora (Connor, Ulla & Upton, 2004, p. 2). Researching specialised corpora has been able to marry up the disparity between linguistic analysis and the situational or cultural context within which such language occurs (Flowerdew, 2004, p.16). This aspect will be discussed in more detail within the “Corpus analysis” section of this chapter, where the implementation of an analytical framework combining a situational, linguistic and functional analysis will be delineated. This framework was chosen to mitigate the limitation of decontextualisation of data frequently connected to corpus research.

A specialised corpus is a corpus that, *per definitionem*, has been compiled with a specific research question in mind to represent a specific register (Baker et al., 2006; Xiao, R., 2010). The application of such specialised corpora for pedagogic usage in particular has seen a vast increase over the past decades, such as within the field of English for Specific Purposes (Belcher 2006; Flowerdew, forthcoming). Studies of this nature frequently compile corpora of less than 1 million words, and the overarching design consideration in the compilation of such corpora favours the inclusion of texts of a particular register or situation over the “representativeness of language across a large number of communicative purposes” (Connor, U. & Upton, 2004, p. 2)

Specialised corpora bear the additional advantage of allowing the data to be manageable enough to be considered context, thus enabling the researcher to overcome or mitigate the criticism of decontextualisation of language from its context. Due to the smaller size of specialised corpora, often not exceeding 1 million words, it is possible to analyse the data quantitatively as well as qualitatively, enabling the researcher to consider language at word, sentence, text level and beyond. Such studies consider, for example, all occurrences of one item in KWIC concordances or consider extralinguistic contextual features of texts included in the corpus, thus being able to overcome the limitations of a purely metrics-driven quantitative approach, and providing insight into language use at two levels: the vertical and horizontal dimension, to use Sinclair’s terminology, or, using Flowerdew’s terminology with its focus on the approach to textual analsyis, the “combin[ation of] top-down and bottom-up” (Flowerdew, L., 2009, p. 396).

Yet, the key limitation in terms of representativeness upholds, especially in small specialised corpus research. It refers to the inability of a corpus to include all examples of language occurring in the specialised context within a corpus (cf. Widdowson, 1998), thus, a corpus will not be able to, in the words of Sinclair, “capture all the patterns of the language, nor represent them in precisely the correct proportions” (Sinclair, 2005, section 2). Such criticism, in line with McEnery and Wilson (2001), can only be countered through careful consideration of ways to achieve representativeness of a corpus.

Representativeness in specialised corpora has been suggested to be measured through the notion of closure (McEnery & Wilson, 2001, p. 166). Closure as a factor of representative has been argued to be the main contributing factor to measure representativeness in specialised corpora (Teubert, 1999), and has been further developed into a framework for the calculation of lexical closure of a corpus by McEnery and Wilson (2001, p. 166).

3.4 Corpus linguistics and written registers

Over the past decades, there has been an ever-increasing body of research using corpus linguistics tools to investigate written texts, including accounts of features of academic writing (Biber & Gray, 2011; Biber & Conrad, 2013; Flowerdew, J., 2017; Flowerdew, 2004; Nesi, Matheson, & Basturkmen, 2017), particularly focused on identifying variation in registers. Particularly influential within the study of genre or register variation has been the work by Douglas Biber et al. (Biber, 1988; Biber, 1995; Biber, Conrad, & Reppen, 1998; Biber et al., 1999; Biber, 2006; Biber & Conrad, 2013).

3.4.1 Register and genre analysis

The distinction between register and genre analyses in this thesis is based on Biber and Conrad's framework for register analysis (Biber & Conrad, 2013). Unlike Biber's earlier work, which used the terms interchangeably (Biber, 1988; Biber et al., 1999; McEnery, Xiao, & Tono, 2006), a distinction is made in the 2013 publication with regarding the analysis of both, register and genre. Here, register

> combines an analysis of linguistic characteristics that are common in a text variety with analysis of the situation of the use of the variety. The underlying assumption of the register perspective is that core linguistic features like pronouns and verbs are functional, and, as a result, particular features are commonly used in association with the communicative purposes and situational context of texts (Biber & Conrad, 2013, p. 2).

Such analytical focus compares with the analysis of genre, in that the genre perspective relies on the inclusion of full texts as it includes structural conventions, for example the use of linguistic features within particular parts of a genre:

> Genre includes [the] description of the purposes and situational context of a text variety, but its linguistic analysis contrasts with the register perspective by focusing on the conventional structures used to construct a complete text within the variety, for example, the conventional way in which a letter begins and ends (Biber & Conrad, 2013, p. 2).

Register analysis, following the framework of Biber & Conrad (2013) considers pervasive ("common") linguistic features, their functional and situational analysis, as well as conventional structures at text level (Biber & Conrad, 2013):

> Registers are described for their typical lexical and grammatical characteristics: their linguistic features. But registers are also described for their situational contexts, for example whether they are produced in speech or writing, whether they are interactive, and what their primary communicative purposes are. One of the central arguments […] is that linguistic features are always functional when considered from a register perspective (p. 6).

Particularly the third aspect of this analytic framework will be useful in bringing in contextual information about the target culture of the texts to help understand the function of the linguistic features within the corpus that often move beyond basic communicative functions. Such functional analysis will also contain information about the way texts are processed within the target community.

3.4.2 Multi-Dimensional (MD) analysis

Biber's (1988) introduction of the multi-dimensional (MD) analysis has laid much of the groundwork for recent studies in the field of register and genre analyses: The classification of both spoken and written registers and sub-registers through the lens of six different dimensions was achieved through analysis of the Lancaster-Oslo/Bergen (LOB) and London-Lund corpora a large general well-balanced and POS-tagged corpora. This seminal work, based on the principal of factor analysis, provided the basis for the identification of variation of written and spoken texts through the linguistic features associated with the different dimensions, and has thus enabled researchers to align their findings along such classifications. Studies applying such MD analysis have been published in a edited book (Conrad & Biber, 2001), and have provided historic and diachronic accounts of registers, well as insights into specialised language, such as scientific or medical texts, just to name a few. Further follow-up work included investigations into academic English, such as an investigation into spoken and written university registers (Biber, 2006), and more recently the language used on the Internet (Biber & Egbert, 2016).

This body of work which applied MD analysis is considered have made the most significant contribution in enabling other studies to align their findings in comparison to other genres or registers. Although such approach will not be replicated in the present study, the findings from the present study, where applicable, will be recontextualised to the six dimensions of the MD analysis, and thus enable shastras to be aligned with other textual registers or genres by means of features identified within the corpus. This is deemed appropriate as "any linguistic feature that can be interpreted as having functional associations" can theoretically be included in MD analyses, including

- Lexical features, such as type-token ratio
- Semantic features relating to lexical classes
- Grammatical feature classes, such as […] personal pronouns; and
- syntactic features […] (McEnery et al., 2006)

Despite not carrying out MD analysis, the “soft application” of the MD analysis by means of using identified frequent linguistic features and evaluating their functionality against specific dimensions within the MD analysis framework has been applied particularly within smaller corpus studies.

3.4.2 Situational, linguistic and functional analyses

A further useful framework to the analysis of registers has been provided by Biber and Conrad (2013), which, unlike the MD analysis, can be applied to any kind of linguistic data to conduct register analyses, regardless of the size, balance and markup of the corpus: “any text sample of any type can be analyzed” (Biber & Conrad, 2013, p. 2). The framework has been designed to support the analysis of registers, genres and styles, and is comprised of analytical approaches to conduct (1) situational analysis, (2) linguistic analysis, and (3) functional analysis. Situational analysis here denotes the analysis of extralinguistic features, such as the context of the text, and can draw on the researcher’s own experiences or observations with the target register, draw on expert informants, previous research or texts from the register under investigation themselves. Linguistic analysis involves the analysis of multiple texts from the same register to indicate pervasive language features of said register. The functional analysis involves the investigation of the link between the situational and linguistic analysis.

It must be added here that the MD analysis and the framework of situational, linguistic and functional analysis are not mutually exclusive but instead, the MD analysis can be used as part of the framework, and has been utilised in this way by Biber (2006) in his analysis of academic registers.

A further significant contribution of the corpus approach to the understanding of features of writing, and positioning them within the registers conversation, fiction, news and academic prose, has been the creation of the Longman Grammar of

Spoken and Written English (LGSWE) (Biber et al., 1999), which draws on both corpus-based frequency data and situational characteristics of registers, and is widely used to enable researchers to correlate functional use of linguistic features with different registers.

3.4.3 Researching academic writing

The focus on researching academic writing may perhaps be surprising here but the reason for drawing on this research is that this thesis, and the corpus compilation, has originally had a pedagogic focus, aiming to establish English for Buddhist Purposes (see 1.2 Background of the study). Furthermore, initial analysis of one shastra as part of a pilot indicated in its wordlist that the register appeared much aligned with features of academic writing. It is for this reason that academic written registers have featured in the comparison of results within this thesis.

Much of the corpus research into academic writing, in addition to the studies on register variation highlighted above, have been driven by an application in teaching (Flowerdew, 2017; Flowerdew, 2004; Flowerdew, L., 2012; Frankenburg-Garcia, Flowerdew, & Aston, 2011; Nesi et al., 2017; Swales, 2006). As such, their research into general academic writing or in specialised domains has been provided to inform the creation of learning materials to help students apply and adhere to common genre conventions (Aijmer, 2009; Boulton, Carter-Thomas, & Rowley-Jolivet, 2012a; Boulton, Carter-Thomas, & Rowley-Jolivet, 2012b; Breyer, 2011; Campoy Cubillo, Bellés Fortuño, & Gea Valor, M. Llui sa, 2010; Gavioli, 2005; Harris & Jae\n, 2010; Lombardo, 2009; O'Keeffe, McCarthy, & Carter, 2007; Partington, 1998; Sinclair, John McHardy, 2004; Tribble, 2000; Weir & Ishikawa, 2010). Though some of these studies have relied on existing and readily available corpora, such as the British National Corpus (O'Keeffe et al., 2007), there has been an increasing number of studies where investigations into academic registers or genres have involved the compilation of specialised corpora to suit their specific purposes (Bhatia, V., Hernández, & Pérez-Paredes, 2011; Bowker, L., 2000; Bowker, Lynne & Pearson, 2002; Flowerdew, J., 2015; Flowerdew, 2017; Flowerdew, 2004; Römer & Schulze, 2010). The next section of this literature review will present the research around corpus-based materials design in more detail.

3.5 Implications of corpus research: language teaching

A key focus of English language teaching (ELT) is to help learners develop fluency. The identification of what makes a language sound “natural” or “idiomatic” is of central importance. Corpus research has made a major contribution by challenging “language ‘rules’ and models presented in teaching publications” based on contrasting data found in language corpora (McCarten & McCarthy, 2010, p. 13). The use of corpora thus bears obvious pedagogic affordances, for example in the selection of typical lexical items and collocations in the language. They furthermore “provide us with large amounts of natural language examples” (Römer, 2006, p. 124). Despite such affordances, it has been observed that there is yet an apparent “resistance” in the use of corpora for pedagogic uses (Römer, 2006, p. 124). Whilst the affordances may be apparent to some, it has been argued by Timmis (2011) that the benefits may perhaps not be as obvious as noted by Römer (2006). The main decision-making factor in language teaching materials design seems to remain reliant on intuition and based on anecdotal evidence (Biber & Conrad, 2010). A reason for this resistance, it has been speculated, may be the lack of user-friendliness of corpus tools, creating a barrier to researchers wishing to utilise corpus data (Römer, 2006). Whilst a general reluctance has been observed among classroom teachers to draw from corpus data in their materials design, the ELT publication sector has seen a gradual shift, departing from being largely based on authors’ intuition to an ever-increasing number of publications based on authentic language use (Ghadessy, Henry, Roseberry, & Sinclair, 2001). The past decades have furthermore seen an increasing effort by corpus linguists to provide guidance and resources to enable language educators to utilise corpora in their practice. Such resources equip ELT professionals with step-by-step guidance of how to investigate corpora and apply their findings in the classroom (Frankenburg-Garcia et al., 2011; Jaen, Serrano, & Calzada Perez, 2010; The Education University of Hong Kong, 2022; Timmis, 2015).

Reppen (2011) distinguishes between three approaches to incorporating corpus data into the classroom: (1) Corpus data obtained through corpus searches conducted and pre-selected by teachers for in-class use; (2) In-class corpus searches of userfriendly online corpora conducted by students; (3) teacher creation of (specialist or learner) corpora or introducing students to existing corpora for exploration by students. These three approaches should not be considered as mutually exclusive but instead can be used in combination to enhance student learning through exploration of and familiarisation with authentic language use:

> Each has certain advantages and, of course, the ideas can be used in combination. For example, a teacher might bring in some prepared concordance lines (i.e. samples of the use of a particular language feature) to introduce new vocabulary, and then later have students search an online corpus to see more examples of the words in context, in order to provide students with greater exposure to the different senses of the target word. (Reppen, 2011, p. 36)

Corpora in their pedagogic application have the affordance of raising students' awareness by noticing patterns of natural language use and contrasting such with their pre-existing knowledge or beliefs of language use (Timmis, 2013). This classroom use of corpus data is underpinned by the pedagogy of data-driven learning (DDL), which will require the learner to explore naturally occurring language, for example concordances extracted from a corpus, to identify patterns (Johns, 2002). This identification of patterns is scaffolded through guided discovery tasks to enhance task achievement. In this way, using corpora in language teaching through the pedagogy of DDL has a focus on awareness raising rather than developing productive language skills (Kettemann & Marko, 2016; Timmis, 2013).

In the situational context of Tibetan Buddhism and monastic education, awareness raising of the language used is a key target of the learning materials to aid understanding of Buddhist English usage. Shastras, corpus text that have been included do not provide models for text production. Students of Tibetan Buddhism are not expected to produce Buddhist philosophies. Instead, much of the language use will be based on memorisation of original sources (see chapter 5.4.3 on memorisation and debate) and as such an enhanced understanding of language used within the context through awareness raising activities is entirely appropriate for the educational context.

For a large part, English language teaching in monastic education in the East sits within the responsibility of native speaking volunteers, often visiting as part of a gap year with no (*Monastery volunteering programme nepal.*2022; *Teaching english to buddhist monks.*2022) or little (van Auken, 2019) formal training in English language teaching required to undertake the role. Where training is offered as part of the training programme, most prominently short online TEFL courses, the qualifications do not cover corpus linguistics. ELT in the context of monastic Buddhist education in the East is predominantly limited to the inclusion of General English language, with a focus on improving basic conversation skills and pronunciation (*Monastery volunteering programme nepal.*2022; *Teaching english to buddhist monks.*2022; van Auken, 2019), at large omitting the specific purpose of the context within which the monks operate. Some rare initiatives have arisen for more structured language instruction with permanent teaching staff, such as at the Dzongsar Khyentse Chökyi Lodrö Institute, India, where a structured 3-year English language curriculum runs alongside the monastic curriculum through *Dharma English* classes,[15] aiming to enhance the number of translators:

> A major aim of these language courses is to train qualified translators, so that the Dharma can be made available to people in their own languages—the fundamental condition for the Buddhadharma to take root and flourish in any country (*Language courses.*2022, para. 10)

The core language programme, however, solely focuses on the development of general English language skills. Instructional materials are limited to the Cambridge University Press *Language in Use* series and Murphy's *Grammar in Use* series and is summatively assessed after the 3-year-period using the standardised IELTS exam (Khashor, 2012). As such, the curriculum excludes the specific purpose. The programme first introduces students to English language use in the Buddhist context through a 2-year internship at an English language Dharma Centre before studying *Dharma English*. As such, the full curriculum does not cater for the specific purpose. Given that no publicly available Buddhist English corpora currently exist, and the compilation thereof is time-consuming. Thus, the lack of corpus-informed materials

[15] The language curriculum is not restricted to English and Chinese is also offered as an optional language

within the Buddhist context is unsurprising as English language instruction sits very much on the fringes of the monastic curriculum, where such a curriculum exists. The creation of language learning materials, catering for the specific context of Buddhist English based on corpus data, will thus make a valuable contribution to existing TEFL practice.

In the context of Tibetan Buddhist English, no previous work has been carried out into the authentic language use within its context. Correspondingly, there are no corpus-based educational materials available. To help bridge this gap, this thesis will exemplify how the linguistic insights gained from the small specialised corpus of Tibetan Buddhist English may be applied in ELT in monastic education. Illustrative examples will indicate how the corpus data of this present study can inform materials design in the context of Buddhist English, with a focus on awareness raising activities to aid the understanding and development of Buddhist English for learners, see Chapter 8.3 Pedagogic Implications.

3.6 Contribution to existing research

The present study aims to make a significant contribution to the body of research. By taking a corpus approach, this study will investigate the written subregister of *shastra* to carry out a first data-driven account of the language used in Tibetan Buddhist English.

The present study will position itself among other investigations into specialised genres and register, and as such aim to position the written subregister of *shastra* within the wider research. As no corpora are currently in existence that would allow such investigation, the compilation of a corpus to represent the written subregister of *shastra* as it is used within the context of Tibetan Buddhism of the Karma Kagyu lineage will be useful. Thus, the present study will contribute to the body of research within this discipline by providing empirical evidence of the linguistic features of said register. Methodologically, contributions will be made to existing research by building a small specialised corpus and thus evaluating previous research into the compilation[16] of corpora (Bhatia, V. K., Hernandez, & Perez-Paredes, 2011; Bowker & Pearson, 2002) and by testing the analytical framework of situational, linguistic and functional analysis (Biber & Conrad, 2013) for the description and classification of a written subregister that is new and emerging in the English language.

3.7 Chapter summary

This chapter has provided an insight into the relevant bodies of research to establish a gap and justify the need for the present study. Research into English language use within the context of Tibetan Buddhism, has highlighted the limitations in the existing body of literature within this field in terms of number, as well as in terms of the methodology applied. A first account of an analysis into the language used within Tibetan Buddhism (Griffiths, 1981) highlighted the issue of language use that proves challenging to comprehend for members outside the target culture, mostly due to the highly frequent use of Sanskrit loanwords, but also the lack of standardisation within the language. Empirically, the present study will be the first of its kind to provide a corpus-based data-driven account of the written subregister *Shastra*.

This chapter further identified research utilising MD analysis (Biber, 1988) as well as the framework of situational, linguistic and functional analysis (Biber & Conrad, 2013) for purposes of classification of register and genre studies, and to enable the positioning of *shastras* based on its linguistic features. Other studies that carried out corpus research into written genres and registers, and more specifically into written academic subgenres using specialised corpora have been identified and will enable the researcher to position the written subregister of *shastra* among general written registers and academic written registers more specifically.

The final section of this chapter explored the literature around approaches to apply such findings to the development of learning materials in the context of English for Specific Purposes, to allow for a practical application of the findings of this study.

[16] A further discussion of the affordances and drawbacks and general application of specialised corpora over general corpora has been provided in the methodology chapter.

CHAPTER 4: METHODOLOGY

4.1 Introduction to the chapter

This chapter will outline theoretical considerations of the specialised corpus research and subsequently, how such theoretical considerations have been implemented in the research design of the present study. Limitations and further developments of the research design will be discussed within the different subsections of this chapter.

4.2 Aims of this study

The main aim of this study is to investigate the language used in the context of Tibetan Buddhism in order to identify the key features of this written subregister to position the register among other general written and academic subregisters.

4.2.1 Research questions

The following main research questions will be addressed in this study:

1. What are the pervasive[17] linguistic features of the written subregister shastra in Tibetan Buddhist English?
2. In relation to question 1, what are the characteristics of such linguistic features?
3. What is the link between such linguistic features and their situational context of Tibetan Buddhist shastras?
4. How do the linguistic features of Tibetan Buddhist shastras compare to other written registers?

4.3 Justification of the specialised corpus approach

The investigation is carried out by taking a corpus approach to the selection and analysis of data. This approach has been chosen as no previous applied linguistics research on the language use in Tibetan Buddhist shastras has been carried out. Thus, this study cannot rely on previous findings to inform the investigation. Taking

[17] The term “pervasive” here is based on Biber and Conrad’s framework for register analysis, which requires “the identification of the *pervasive* linguistic features” (Biber & Conrad, 2013, p. 6)

an empirical approach allows the researcher to obtain examples of language in use. Furthermore, utilising means of corpus linguistics methods will help the researcher identify pervasive linguistic features in the written subregister of shastra (research question 1) in order to investigate their characteristics (research question 2). Providing insight into the situational context within which such features occur will (research question 3) allow the researcher to illustrate the suitability of the linguistic features in light of the purpose of the register. Further comparison of the data with larger written general and academic registers allows the researcher to position the written subregister of shastra within the wider discipline (research question 4).

4.3.1 Naturally occurring language vs introspection

One of the advantages of the corpus approach to data analysis is the ability of the researcher to investigate language as it naturally occurs, also referred to as attested data. Given that there is only limited research conducted into the language used within a Buddhist context, and no research conducted from a linguistic perspective into the written subregister of *Shastra*, the investigation into naturally occurring data is the only appropriate approach to gain insights into language use within the specialised context.

This affordance is deemed critical for the present study as, firstly, there is insufficient data available to allow examples to support hypotheses to be drawn purely from researchers' introspection (Fillmore, 1992). Secondly, this approach enhances the reliability of the study as the approach allows replicability, and its findings are based on empirical research. The limitations of introspection have further been highlighted in that there may be discrepancies between researchers' *intuition* about language use and *actual* language use (cf. Biber & Gray, 2011). Taking an empirical approach to language analysis will mitigate the limitations of researcher bias which are the main point of criticism where introspection informs language analysis.

4.3.2 Limitations of the specialised corpus approach

It is, however, unavoidable for researcher bias to influence the analysis of the data extracted from the corpus. “One of the principle uses of a corpus is to identify what is central and typical in the language” (Sinclair, J., 1991, p. 17). Albeit the initial analysis being conducted by means of computational calculations of word lists and keyword lists to support the researcher in the identification of what is deemed *central* and *typical* for the specialised language investigated in the present study, the decision of which items to investigate in more depth, even if mitigated by consideration of frequency, are taken by the researcher and are thus subject to researcher preference.

4.4 Corpus design

The following section will outline at its outset the purpose and intended applications of the corpus and the way in which the use of the corpus has informed the decision-making process regarding the corpus design. This section will additionally provide an overview of the corpus architecture and highlight the issues and limitations of the data.

4.4.1 Process of corpus design

The design of a corpus is influenced by two factors (McEnery & Wilson, 1996): practical and theoretical factors. Practical considerations, such as time and resource constraints, have affected the design of the corpus in that it had to be feasible to be implemented by one researcher within a limited period to form a PhD thesis.

At a theoretical level, the design is determined by the purpose and use of the corpus, based on the research questions to be answered. The impact the purpose of a corpus and the underlying research questions have on the design of a corpus have been widely asserted in the scholarship (McEnery & Hardie, 2012; Xiao, 2010). In the context of Language for Specific Purposes (LSP), “corpora are built to answer research questions related to teaching or lexicographic projects" (Williams, 2002, p. 46). The aim of the present study is to investigate the language studied by Tibetan Buddhists through identification and analysis of salient features of the written subregister of *shastra.*[18] Such investigation was initially pedagogically motivated with an intended application of the findings in materials development for the field of English for Specific Purposes (ESP) in the context of Buddhism. The present study, carried out through the compilation and analysis of a corpus, thus follows a pedagogical purpose, and such purpose has heavily impacted on the corpus design.[19]

Conventionally, a research inquiry has at its outset clearly defined research questions which subsequently inform the justification of the methodology employed in the study. The present project deviated from such an approach.

As the present study is the first investigation to consider the genre of *shastra* in its English translations through a linguistic lens, previous related literature on the subject matter could not inform the present study or the formulation of hypotheses to be tested beyond the linguistic investigations that have been conducted within the field of translation studies or Tibetology. Such studies, however, have not considered the body of text but dealt with the language under investigation at a single text level or through analysis of multiple translations of single source texts, thus not provided insight into linguistic considerations beyond the text, often limited to single word-level features.

It is for this reason, that the corpus was designed in a way that allowed an iterative approach (see Figure 3 below) between the investigation of the data and the research questions underpinning this project. This allowed the researcher to explore the data at the outset of the investigation and, based on preliminary findings, be in a

[18] *Shastras* are commentaries that are composed to explain a specific topic within Buddhist philosophy. Each *shastra* compiles and comments on relevant writings from Buddhist scriptures on the specific topic. In this sense, a *shastra* is much like a compendium.

[19] It was indicated in section *1.2.1 Changing the focus of the thesis* that the focus of this thesis had to change due to unforeseen external circumstances. At that time, the compilation of the corpus had already taken place

position to formulate more specific research questions to narrow the analysis later on.

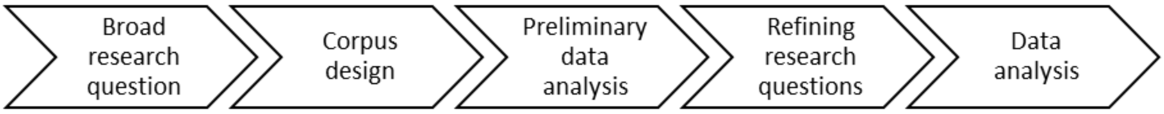


Figure 3: Process of data analysis

This iterative, corpus-driven approach to data analysis required the corpus data to provide sufficient flexibility to cater for a range of emerging research questions that would not limit the investigation to, for example, a genre or register perspective, whilst still being feasible within the time and labour constraints of the project. The subsequent description and justification of the corpus design and approach to data analysis will further illustrate how this was applied to and has impacted on the corpus design and approach to data analysis.

4.4.2 Overview of the design

Bowker and Pearson (Bowker & Pearson, 2002) present a framework for building LSP corpora, considering corpus size, number of texts, medium, subject, text type, authorship, language and publication date. The framework thereby provides a guideline for the considerations made within defining a sampling frame. The sampling frame used in the present study is based on external sampling criteria, as also proposed by Biber (1993) in his work on large corpora.

A frequent limitation of the work conducted using specialised corpora is the lack of consideration of reliability measures. The present study will implement reliability measures by means of measuring the lexical representativeness of the corpus through calculating lexical closure as proposed by McEnery and Wilson (1996). The study will further consider corpus permanence as a measure to evaluate the reliability of the analysis of the present study, aligned with the work by Hunston (2002, p. 30). Such deliberations will create transparency in the process and design of the corpus and thus ensure replicability of the present study, allowing the repetition of the investigation and the testing of the findings of this study by others.

4.4.3 Sample

4.4.3.1 Internal vs external criteria

In corpus design, the research is favourable to the utilisation of external criteria to determine the corpus sample in order to represent the language under investigation.

> The contents of a corpus should be selected without regard for the language they contain but according to their communicative function in the community in which they arise (Sinclair, 2005).

This proposition is furthered by way of considering the use and purpose (Aston & Burnard, 1998) of the corpus, as well as situational occurrence of the language (Xiao, 2010, p. 149) to determine the criteria for text inclusion therein. Although the above references make statements about the compilation of large corpora, they are yet deemed appropriate and well suited to be applied to the present specialised corpus, due to its focus on the specialised context of Tibetan Buddhism, and the pedagogic purpose the corpus follows, with a stronger focus on genre and register features at the lexical level and a lesser focus on grammatical items which would require perhaps a combination of external and internal selection criteria where the interplay between both dimensions of selection criteria are cyclical in nature (Biber, 1993): A corpus, initially compiled based on external criteria would subsequently be investigated for its “homogeneity” (Sinclair, 2005), that is the balance of distribution of linguistic features within the corpus. Such internal criteria may inform, at the second instance, an exclusion of texts from a corpus that were previously selected based on situational characteristics, based on a discord other texts contained within the corpus due to the linguistic features contained within said text. This cyclical process of text selection within corpus design would bare the benefit of providing a measure of representativeness at the linguistic level, in practice frequently with a focus on grammatical items (Williams, 2002).

Such approach is appropriate when building a large corpus. As this study provides a first investigation into the written subregister of the shastra to provide insights into the language used within such texts, such approach would have been beyond the scope of this exploratory study. Applying the iterative process of external and internal criteria to inform the corpus sample cannot be taken for reasons of feasibility and restriction of resources of the present study, as it would have required a comprehensive analysis of the grammatical features within a POS-tagged corpus,[20] as, for example, suggested by Biber (1993).

The issue of homogeneity will yet be considered within the corpus analysis. For a specialised corpus, measuring representativeness with a focus on lexical items by means of lexical closure (McEnery & Wilson, 2001; Xiao, 2010) and internal variation within the corpus by means of keyword lists (Hunston, 2014), is deemed more appropriate. The present study, however, will not consider such measures as part of the *sampling process* but instead determine text selection solely on external situational characteristics as elaborated below, and consider lexical closure of, and internal variation within, the corpus as part of the analysis. The data presented must thus be seen within the limitations resulting from the corpus design considerations. An in-depth analysis of the linguistic features to determine the representatives of each text through internal factors would provide an interesting future research project.

4.4.3.2 Discourse community approach to sampling

Basing the sampling considerations on external criteria is advantageous in that it allows the researcher to determine the criteria and provide a justification for the inclusion or exclusion of texts into the corpus prior to any linguistic analysis being conducted. It thus helps overcome the conundrum associated with internal selection criteria. Internal selection criteria means that naturally occurring language is selected due to the language use within each text. Where only internal criteria are applied, it will likely skew the results of the analysis (Xiao, 2010).

[20] Analysis of grammatical patterns was considered at the outset of the study and tested using automated POS annotation in WMatrix (Rayson, 2009) and TagAnt (Anthony, 2015), based on TreeTagger. The results were limited in accuracy and could have only been overcome through manual POS annotation. This labour-intensive process was deemed beyond the feasibility of the present study. Further study based on a POS-tagged corpus to investigate grammatical patterns is recommended

Application of external criteria to the text selection further mitigates the issue of researcher bias to the selection process at the text level. The determination of the sampling criteria, the sampling frame, is still subject to researcher bias. It is for this reason that the present study employs a discourse community approach, basing the sampling of texts on the discourse community that utilises such language. Such an approach would thus imply that the selection of texts includes all publications on the subject. It is obvious that this approach is too broad to be useful to help determine the final selection of texts, and to reliably represent the target communities' language use. Narrowing down this wide selection of texts using a discourse community approach seems appropriate as it distances the researcher from the selection process and places the discourse community at the heart of the decision-making. Within the target community, a range of different text types could be included such as transcripts of oral textual teachings,[21] prayers, transcripts of conversations or discussions between members of the discourse community on the subject matter, public speeches, etc. Although such an investigation and such a comprehensive corpus is considered valuable and would provide insight into the differences between the different modes and registers within the specialised language, it is not feasible given the resources and time constraints of a PhD thesis. Furthermore, given the intended use of the corpus being of a pedagogical nature, the decision was made to base the textual selection on the set reading list of the traditional curriculum of the Karma Kagyu School of Tibetan Buddhism. These set reading lists are comprised of texts of the Buddhist written subregister of shastra.

> As part of his vast activities in preserving the Kagyu educational tradition, His Holiness the Sixteenth Gyalwa Karmapa identified eight major texts as the focus for the Kagyu educational program. These texts are centerpieces of the distinctive Kagyu educational system, and the texts form the core curriculum for studies in the Kagyu shedra. (*The eight great texts of the kagyu tradition.*n.d.)

[21] These teachings are similar in nature to the academic lecture format, whereby a qualified teacher delivers a literature-based talk on a subtopic within Buddhist philosophical thought

The collection of shastra*s* that are studied by ordained monks or nuns in Buddhist colleges in the East as part of their curriculum are called *shedras*, or, by Western laypeople studying Buddhist philosophy as part of so-called "*Shedras* in the West".

> There [within the *shedra*] they [the students] would be able to pursue studies under the direction of highly qualified masters on the topics of Madhyamaka (The Philosophy of the Middle Way), Paramita (The Path of the Bodhisattva), Pramana (Logic and Epistemology), Vinaya (Monastic Discipline), Abhidharma (Higher Teachings) and Tantra. These topics were taught through an exegesis of the classical philosophical works (shastras) of Indian Buddhism together with the relevant explanatory commentaries composed in Tibet. Such a system allowed for an extraordinary depth of knowledge and insight that could act as a superb foundation for meditation. Furthermore, those who graduated from the shedras would, in many cases, subsequently take up responsibilities as textual and philosophical teachers themselves, thus assuring the continuation of such sacred knowledge (Dechen, 2008)

A corpus of a written subregister of *shastra* is well-placed to be representative of the language used within Buddhism, lexically speaking, in that it is functionally similar to the genre of compendium: A *shastra* is a commentary focused on a specific subtopic of Buddhist philosophy, drawing together the relevant Buddhist literature on said subject matter. The canonical philosophical texts (*shastras*) included within the curriculum have been selected to provide the student of such texts with a detailed understanding of Buddhist philosophy. Therefore, such genre, and textual selection within this genre, is well suited to cover the range of topics contained within it. Basing the sampling frame on the selection of *shastras* that comprise the curriculum for Buddhist philosophical thought as taught through the *shedra* curriculum of the Karma Kagyu School of Tibetan Buddhism will thus provide topical coverage and correspondingly, ensure that the corpus is lexically representative of the naturally occurring language. Simultaneously, this discourse community approach to sampling allows the researcher to distance herself from the sampling process, thereby minimising researcher bias in the selection of the texts that comprise the corpus. The sampling decision is made by the community in which texts are studied, and in which the corpus will find application.

Two issues in the text selection had to be overcome at this point. Firstly, colleges for the study of Tibetan Buddhist thought, held in the East, use Tibetan as the language of instruction, and the texts studied are composed in classical Tibetan. Secondly, much akin to the study of, for example, English literature, texts that are included and studied within such *shedras* are based on a canon, similar to the principles of a canon in Western literary education. This implies that, although there is a major overlap in key texts studied, there may be some deviations regarding the set texts between different *shedra* curricula, and between the different translations that are selected for study as part of *shedras* taught in the English language.

The present study has been based on the Mikyo Dorje *shedra* curriculum as taught by the Karmapa International Buddhist Institute (KIBI) in New Delhi, India, and within the Dechen community of Tibetan Buddhism in the West. These communities have been selected as they are in the unique position of following the same curriculum, with the KIBI teaching texts in their original language, and the Dechen community utilising the texts in their English translation for their "*Shedra* in the West". Thus, the present study investigates the naturally occurring language use of texts in their English translation as selected by the discourse community. One exception has been made regarding the inclusion of texts in the corpus, which an unpublished working draft of a translation, disseminated amongst the community members of the Dechen sangha, utilised as part of their *shedra* teachings: "The lamp of excellent discrimination of the system of Zhentong Madhyamaka".

Such considerations resulted in the following list of texts (see Table 1 overleaf) to comprise the corpus is based on the Mikyo Dorje Shedra curriculum of the Dechen community (Dechen, 2008).

Corpus text	Title	Author	Translator	First published	Year of edition	Place of publication	Publisher
1a	Gateway to Knowledge Vol. I	Jamgon Mipham Rinpoche	Erik Pema Kunsang	1984	1997	Hong Kong	Rangjung Yeshe Publications
1b	Gateway to Knowledge Vol. II	Jamgon Mipham Rinpoche	Erik Pema Kunsang	1984	2000	Hong Kong	Rangjung Yeshe Publications
2	The Jewel Ornament of Liberation	Gampopa	Khenpo Konchog Gyaltsen Rinpoche	1998	1998	Ithaca, New York, US	Snow Lion Publications
3	The Instructions of Gampopa: A Precious Garland of the Supreme Path	Gampopa	Konchok Rigzen	2010	2010	Taipei, Taiwan	Central Institute of Buddhist Studies Leh-Ladakh
4	Clarifying the Thought of Rangjung	Jamgon Kongtrul Lodro	Jampa Thaye	1996	2015	Bristol, UK	Ganesha Press
5	The Lamp that Dispels Darkness: A Commentary on Karmapa Rangjung Dorje's 'Distinguishing Consciousness and Primordial Wisdom'	Karma Thinley Rinpoche	Adrian O'Sullivan	2013	2018	Bristol, UK	Ganesha Press
6	The lamp of excellent discrimination of the system of Zhentong Madhyamaka	Karmapa Mikyo Dorje	Jampa Thaye	1998	1998	Bristol, UK	Ganesha Press
7	Buddha Nature: The Mahayana Uttaratantra Shastra with Commentary	Arya Maitreya	Rosemarie Fuchs	2000	2000	Ithaca, New York, US	Snow Lion Publications

Table 1: Publication details of the Mikyo Dorje Shedra Tibetan Buddhist Corpus (MDSTB corpus)

4.4.3.3 Corpus architecture

The following description of the Corpus of Tibetan Buddhist English follows the categories proposed by Bowker and Pearson (2002), and its delineation serves the purpose of creating replicability of the present study, as illustrated in Table 2 below:

Mikyo Dorje Shedra Tibetan Buddhist Corpus (MDSTB corpus)	
size	281,290
number of texts	8 (full text length)
medium	Written publications
mode of delivery	Texts are read out (listenend to) and read (studied)[22]
subject	Buddhist philosophy as taught as part of the Shedra curriculum of Tibetan Buddhism, Karma Kagyu lineage
text type	Shastra
language	English (in translation from commentaries written in classical Tibetan, collating and explaining information from Sanskrit root texts)
publication date	1997-2018 (based on first editions of 1984-2013)

Table 2: Corpus architecture of the MDSTB corpus

4.4.3.4 Size, number of texts, text length

The literature on corpus compilation in general and specialised corpora specifically is in agreement that there is no set minimum size that is required for a corpus study. The number of texts and words per text required is informed by the research inquiry (Flowerdew, 2004; Sinclair, 2005; Xiao, 2010) as well as practical considerations (Flowerdew, 2004, p. 18). Basing the sampling of the corpus texts on the curriculum for an exploratory study with the focus on qualitative inquiry, a small size of the corpus (281,290 words) comprised of 8 corpus texts of varying lengths, as indicated in Table 3, is deemed appropriate as the main focus of the analysis will be considering lexical rather than grammatical features. Lexical representativeness of

[22] Further detail on the transmission and study of *shastras* has been provided in chapter 7.4 *Shedras* in the West, explaining how texts are studied within their target community, and in chapter 5.4.3 The practice of memorisation and debate in Buddhism, explaining how Buddhist texts are studied and memorised in order to inform the Buddhist practice of debate

the corpus will be investigated later in this chapter. Furthermore, keeping the corpus small in size will allow the researcher to investigate and consider contextual features of the corpus to gain an understanding of the potential causes for the linguistic phenomena observed therein. The corpus size has been highlighted in Table 3 below, broken down into size of the different texts contained within the corpus:

Text	1a	1b	2	3	4	5	6	7	corpus
Tokens	19,882	23,565	68,860	23,746	13,640	28,383	7,216	95,998	281,290

Table 3: Corpus size and text length of the MDSTB corpus

The text lengths, as can be seen in Table 3, vary in length between 7,216 tokens and 95,998 tokens. The decision was made to include whole texts as part of the corpus. Not only has this approach been proposed for large general corpora, as can be seen from Sinclair's statement below:

> Personally I would like to see 'whole text' as a default condition, thus classifying sample corpora as one of the categories of special corpora... To me the use of small samples is just a remnant of the early restraints on corpus building, and the advantages of whole texts can be set out in powerful argument. The use of samples of constant size gains only a spurious air of scientific method, since it confers no benefit on the corpus, and is as practical as Genghis Khan's fabled policy of having all his soldiers the same height (Sinclair, John, 1995, pp. 27-28)

Research into specialised corpora have promoted the same approach. Greater importance is placed on the selection of texts by, for example, inclusion of texts "relevant to the same topic, oriented towards the same purpose, and produced under similar contextual circumstances", favouring full texts (Gesuato, 2011, p. 49). Full texts have been identified to bear the advantage of enabling researchers to consider structural features, as, for example, crucial for register analyses (Connor & Upton, 2004, p. 2), or to consider the distribution of observed features across a text to identify variations and discourse functions (Bowker, L. & Pearson, 2002; Connor & Upton, 2004; Flowerdew, 2004; Hoey, 2007; Upton, 2002).

The present study had, at its core, the approach of the language in use with as little researcher bias as possible, allowing flexibility to investigate the data to provide answers to emerging research questions, as outlined in *Figure 3: Process of data analysis.* Text samples would have prevented an investigation into genre features. Furthermore, as texts progress from subject to subject gradually, it was feared that the inclusion of mere samples would have led the corpus to misrepresent certain topics, as aspects of the logical development of arguments contained within texts would have been omitted. The drawback of this approach is the resulting imbalance in the corpus in terms of varying text lengths. This dichotomy between the insights to be gained through full text inclusion on the one hand and imbalance and thus potentially skewed results generated by the corpus on the other, was identified by Sinclair (2005, section 3 para. 5): “The problem is that long texts in a small corpus could exert an undue influence on the results of queries, and yet it is not good practice to select only part of a complete artefact”.

This dichotomy is considered within the analysis through consideration of distribution across the corpus and a strong focus on qualitative aspects of the observed linguistic phenomena. Additionally, this shortcoming is considered within the limitations of this study. Yet, such limitation does not render the findings of this study meaningless – particularly considering the exploratory nature of this study of a written subregister that has not previously been considered through this lens. It is anticipated that future research will compile a much larger corpus that fulfils a more general purpose of representing Buddhism in general, including other genres, schools and traditions, and such a corpus would cater for a different range of research questions altogether, and therefore be representative of the language used within Buddhism in general. The data that can be obtained from the present corpus, however, does not make such claims, and the claims that *can* be made based on observations within the present corpus are based on the pedagogical purpose it will serve in its application.

4.4.3.5 Subject, text type, language

The sample being based on the curriculum for the study of Buddhist thought ensures an appropriate topical coverage and subject relevance across the corpus. The canonical philosophical texts within the Mikyo Dorje Shedra curriculum provide the student of such texts with a broad understanding of Buddhist philosophy. Therefore, the written subregister of the *shastra* and textual selection is well suited to cover the range of topics contained within it.

Shastras are written commentaries that are composed to explain a specific topic within Buddhist philosophy. Each *shastra* compiles and comments on relevant writings from Buddhist scriptures on a specific subtopic of Buddhist philosophy. In this sense, a *shastra* is much like a compendium. *Shastras* in this corpus are texts in translation – they were originally composed in classical Tibetan. Such Tibetan texts were composed to unpack subject matter within Buddhist philosophy, thus collating, and containing lengthy quotations, from root texts that are then commented on. Such root texts were originally composed in Sanskrit. In some cases, the English translator is to rely on the quality of the original translation from Sanskrit to Tibetan when producing the English translation. In rare cases do translators have sufficient experience and knowledge to consult the Sanskrit root texts for their English translations, which is why the quality of some translations has been disputed (Denwood, 1983). It is for this reason, that the decision-making of the texts was based on the discourse community.

In many ways, the increasing engagement with Buddhist texts in the West in recent years, and the growing number of translations available, alongside an ever-growing expertise of Buddhist philosophy in the West, as well as linguistic competence in classical Tibetan and Sanskrit, would provide a platform for future investigations by means of diachronic corpora to indicate shifts in language use within this specialised context.

4.4.3.6 Authorship

Authorship, as will become clear to the reader after becoming aware of the textual genre as delineated in the section above, is multi-layered. Authorial work of the root text undergoes the process of translation, which in itself adds an authorial level to the text. This is further supplemented by the commentary provided, collating the textual sections of the root texts. Such work is then further translated into English, adding yet another authorial level to the texts.

The description of the sample above was offered to provide transparency on the sampling process, thus ensuring replicability of this study.

4.4.3.7 Corpus permanence

Similarly, a corpus needs to be viewed as a snapshot of language use at a specific time. Even within the short duration of a PhD study, new editions of the texts included in this corpus have been published, leading to potential changes in the language use within the specialised context over time. Both issues, representativeness and permanence will be considered in more detail within the "Corpus design" subsection "Corpus representativeness" of this chapter. Although strategies are in place to optimise the corpus, any corpus does have limitations and this final section has been included to show the researcher's awareness of such.

4.5 Corpus representativeness: lexical closure

"One of the principle uses of a corpus is to identify what is central and typical in the language" (Sinclair, 1991, p. 17). Leech (1992) stated that corpora are designed to represent a language population,[23] Such an aim has been reiterated and expanded on by Sinclair who considers the aim of the corpus designer to "make their corpus as representative as possible of the language from which it is chosen" (Sinclair, 2005). The guiding principle is that the findings based on a corpus should enable the researcher to make general statements of the language or language variety under investigation (cf. Leech, 1991; Xiao, 2010). Biber contributes to such definitions by elaborating on how representativeness can be achieved. From his perspective on general corpora, "representativeness refers to the extent to which a sample includes the full range of variability in a population" (Biber, 1993, p. 243). Any claims made on the basis of corpus research must inevitably be linked to the representativeness of

[23] The term *population* was not used by Leech but has been adapted from Biber (e.g. 1993). Leech refers to the term language or language variety

the corpus to the target language, as any corpus will be limited in the sample of language that it can contain.

Although there seems general consensus in the aim of achieving representativeness in corpus design, the extent to which such aim is achieved differs within the literature (Xiao, 2010). Regarding small specialised corpora, there appears much less focus on achieving representativeness and a number of core texts on specialised corpora make no suggestion as to how this can be achieved (Bowker & Pearson, 2002b). A reason for this could be the way in which data is collected for investigations into specialised language, in what McEnery and Hardie (2012) label "opportunistic corpora" – where the researcher makes use of data they have available, "because it is there", such as, for example, where a collection of essays from a group of students is analysed. Representativeness in specialised corpora has been suggested to be measured through the notion of closure (McEnery & Wilson, 2001, p. 166). Closure has been argued to be the main contributing factor to measure representativeness in specialised corpora (Teubert, 1999). A framework for the calculation of lexical closure of a corpus has been provided by McEnery and Wilson (McEnery & Wilson, 2001, p. 166), and has been adapted in the present study (chapter 4: word frequency analysis).

4.6 Data analysis framework

4.6.1 Keyword approach

The linguistic analysis as part of the register analysis approach described above relies on the "identification of pervasive linguistic features in the variety" (Biber & Conrad, 2013). The identification of such features has been informed by means of word frequency lists, keywordlists and keykeywordlist. The keyword approach has also been applied in other studies, such as in Xiao and McEnery (2005) who conducted a genre analysis combining Tribble's keyword analysis (Tribble, 2000) with Biber's multidimensional (MD) analysis (Biber, 1988). The present study, however, did not utilise the tool Wordsmith Tools to conduct such analysis, but instead used the tool AntConc (Anthony, 2004) for such analysis. Further analysis of keykeywords were conducted manually. Keyword and keykeyword lists relied on the use of a reference corpus, and thus simultaneously afforded the comparison of the present findings with other written registers (through the use of the BNC written subcorpus as a reference corpus).

4.6.2 Comparison of findings

Findings from the linguistic analysis have been compared with findings from other studies, predominantly conducted on general written or academic written registers, such as the work of the Longman Grammar of Spoken and Written English (Biber et al., 1999). This allowed the researcher to align the present corpus alongside other written registers. Although a MD analysis will not be carried out as part of this study, it will yet be possible to align the findings from the present corpus with Biber's dimensions from this study, (Biber, 1988) based on the following features within this corpus, aligned with the dimension 1: "involved versus informational production" as illustrated in Table 4 below:

Dimension 1: Involved versus informational production	**Features: positive (+); negative (-)**	**Application in this thesis**
	(+) Second-person pronouns	Chapter 6: infrequent use of personal pronouns; frequent use of indefinite pronoun *one*
	(+) First person pronouns	
	(+) Indefinite pronouns	
	(-) Type/token ratio	Chapter 5: lexical repetition
	(-) Nouns	Chapter 5: word frequency lists; prepositional premodified noun phrases
	(-) Prepositions	

Table 4: Features of patterns associated with Dimension 1: "Involved versus informational production" and their application in this thesis

CHAPTER 5: LEXICAL CLOSURE, LEXICAL REPETITON AND WORD FREQUENCY LISTS

5.1 Introduction to the chapter

The aim of this chapter is to provide insight into the “aboutness” of the Tibetan Buddhist *shastras* in terms of its lexical closure properties, lexical repetition, and word frequency lists.

Lexical closure will be used to indicate the lexical representativeness of the corpus, and as such, function as a reliability measure. Lexical closure will be calculated by means of lexical growth analysis, and its closure properties will be compared to the general English language corpus BNC as well as the specialised CRAFT corpus. In this way it will be illustrated that, despite the relatively small size of the specialised corpus under investigation in this study, the corpus achieves near closure and is, as such, lexically representative of the language it represents.

The second part of the chapter will measure the lexical repetition of the MDSTB corpus by means of type-token ratio (TTR). It will be indicated that the corpus has an unusually low TTR compared to other written registers, particularly given the small size of the MDSTB. Lexical closure properties, it will be argued, will account to some degree for such findings. The chapter will further signpost to the analysis in chapter 7, where the frequent repetition of headings and subheadings within the corpus texts are indicative of a high lexical repetition.

The final section of this chapter will investigate the language of Tibetan Buddhism through word frequency lists (wordlist, lemmatised lexical wordlist and keykeywordlist). The word frequency list provides a comparison of the 25 most frequent items within the MDSTB corpus compared to the BAWE and BNC written corpora. Such analysis will indicate shared features and differences, and enable the researcher to position the written subregister shastra among other written registers. The BAWE corpus has been selected to compare the MDSTB corpus against a corpus comprised of written registers from a wide range of academic disciplines that is readily and freely available. As such, comparison of findings with the BAWE corpus will provide an indication of similarities and differences to academic written registers, yet, the main caveat is that the BAWE is comprised of unpublished student writing, and the conclusions drawn from such analysis, although they provide useful indications, must be considered with caution.

Finally, linguistic features for further investigation in subsequent chapters will be identified this way.

5.2 Key considerations and terminology

Before presenting the frequency data of the corpus, some underlying considerations need to be addressed, aligned with the propositions by Archer for researchers working with wordlists (Archer, 2009a, pp. 4-5):

1. what counts as a word?
2. what we mean by frequency?
3. why frequency matters so much?
4. the consistency of the various keyword extraction techniques
5. which of the (key)words captured by keyword/word frequency lists are the most relevant (and which are not)
6. whether the (de)selection of keywords introduces some level of bias
7. what counts as a reference corpus and why we need one
8. whether a reference corpus can be bad and still show us something what we gain (in real terms) by applying frequency and keyword techniques to texts

The subsequent sections within this chapter will consider the proposed questions as the preamble of the generation of wordlists, keyword lists and keykeyword lists.

5.2.1 Words, items, tokens and types

A “word” in this study, refers to an orthographical unit that is preceded and followed by a white space. This definition is useful in that it allows individual *words* to be identified and counted by the corpus software. This definition does not discriminate between the meaning or function of the individual word but merely counts orthographic distinct units (pre and succeeded by a white space). Broadly, the literature distinguishes between *lexical* and *grammatical items*. The present study will focus predominantly on *lexical items*.

This is justified as automatic POS annotation of the corpus proved problematic and showed a high frequency of errors in the automated annotation. Two tools were used to attempt POS annotation of the corpus: WMatrix (Rayson, 2009), utilising the CLAWS part-of-speech tagger (Garside, 1987) and TagAnt (Anthony, 2015), the automatic POS annotation tool developed by Laurence Anthony, based on Helmut Schmid's TreeTagger. Accuracy proved particularly problematic with Sanskrit loan words or proper nouns (although some POS tags were accurate due to the syntactic positioning of such items), or with items with unusually highly frequent function, such as the frequent occurrence of *one* in its function as a pronoun – just to mention one example. Manual analysis was considered beyond the scope of this project, particularly given that any findings would be limited in their ability to allow for generalisations to be made as the corpus is too small and imbalanced to make reliable claims of grammatical preferences of the written subregister under investigation. A meaningful analysis of grammatical items would require a significantly larger corpus with higher frequencies of low frequency items (Biber et al., 1998; Conrad & Biber, 2001). Where an investigation into grammatical items was deemed justified and appropriate, such as the case for the keykeyword *one* (see chapter 5), such analysis was easily conducted through the advanced search function in AntConc (2004; 2019), using a list of the closed group of this grammatical category.

This study also refers to *words* as *tokens* or *types*. In terms of frequency, in a corpus all items within the corpus are the number of *tokens*. By contrast, types are the number of unique items within the corpus. As such, any item that is repeated within the corpus will appear as 1 unique type, whereas its count of occurrences will make up its number of tokens.

5.2.2 Frequency and frequency lists

Frequency in this study refers to the count of occurrences of items within the corpus. Word frequencies are obtained by means of wordlist generation. In this sense, a wordlist in this study refers to a list of words, each counted for its occurrence, contained within the corpus, sorted in descending order by frequency.[24] Frequency has particularly been set at the core of corpus studies of a pedagogic nature, indicating that the more frequent a word occurs, the more important it is to be taught (Flowerdew, 2015; Leech, 2011). The generation of word frequency lists has thus been widely proposed as the starting points for linguistic analyses, not limited to studies of a pedagogic nature (Archer, 2009b; Baron, Rayson, & Archer, 2009; Bowker, 2000; Bowker & Pearson, 2002; Leech, 2011; Tribble & Jones, 1997).

The present study utilises three types of word lists:

1) A word frequency list
This wordlist was generated to allow for the data to be compared to other wordlists, such as the BNC written corpus and the BAWE corpus.[25]

2) A word frequency list of lexical items
A lemmatised word list summarises different word forms that are the result of inflection, providing frequency counts per node (base form) as the basis for the word frequency counts, yet displaying the breakdown thereof into the different word forms. The BNC lemmalist was used to generate the lemmatised wordlist, which has been supplemented with specialised terminology from the MDSTB corpus as further explained in chapter 6. A stoplist was used to limit the items in the wordlist to lexical items only, thus focusing on the topical makeup of the corpus rather than the grammatical aspects.

3) Keykeywordlist
This will be elaborated on within the next section

The wordlist and wordlist of lexical items in this study were generated using the corpus software AntConc (Anthony, 2019). The wordlists generated provided *observed* frequency counts. Subsequent calculations were carried out semi-

[24] Other means of sorting are also possible in the Corpus tool AntConc (Anthony, 2019), such as by word beginning or word ending

[25] The BAWE corpus has been selected to compare the MDSTB corpus against a corpus comprised of written registers from a wide range of academic disciplines that is readily and freely available. As such, comparison of findings with the BAWE corpus has provided an indication of similarities and differences to academic written registers, yet, the main caveat is that the BAWE is comprised of unpublished student writing

automated and manually,[26] for example, to determine *relative* frequencies (per million tokens), to allow for meaningful comparisons of the data.

5.2.3 Keywords and keykeywords

Keywords refer to tokens within a corpus that are over- or underused (positive or negative keywords) when the word frequency list of said corpus is compared to the word frequencies of a reference corpus. In this sense, what constitutes a keyword differs depending on the reference corpus selected, placing great importance on the selection of an appropriate reference corpus (Archer, 2009).

The present study utilises the BNC as a reference corpus, more specifically its written subcorpus. One reason for this choice is that the corpus satisfies the criteria of size; furthermore, a large volume of research has been carried out on this corpus, allowing the present study to be compared to the findings from the BNC, for example in the measures of lexical closure. Although the publications within the MDSTB corpus are, on average, from the late 1990s, early 2000s, the texts in translation are historic, which will inevitably have an impact on the language used therein. Thus, it is deemed appropriate to have used the older yet freely publicly available BNC as a reference corpus. The BNC will be able to help extract keywords within the MDSTB corpus that indicate “aboutness” on the one hand, and stylistic and register specific features on the other. The written subcorpus was selected to avoid the extraction of features that are specific to written texts over spoken language.

One caveat of the extraction of keywords is that a keywordlist may yield items that are highly frequent in a small number of texts within a corpus, yet not unusually highly frequent in other texts. This is a crucial point given the architecture of the present corpus with an imbalance of text lengths, a small number of texts comprising the corpus (only 8 in number) and a small corpus overall (approx. 280,000 words). Thus the calculation of keykeywords is important to help detect only those keywords that are key in all texts of the corpus, in other words, keykeywords can provide

[26] Data was exported from AntConc as .txt files, which were subsequently imported into Microsoft Excel for further calculations or analyses to be carried out.

insights into the “dispersion of a key word in the corpus” (Baron et al., 2009). Scott, who introduced the notion of keykeywords, specifies that keykeywords are “words which are key in a *large number* of texts of a given type” (Scott, 1997, p. 237). In the present study, given the small number of texts contained within the corpus, it is deemed more appropriate to narrow this down to include only keywords that are key in *all* texts of the corpus, thus only extracting words that are overused within the specialised language context and register.

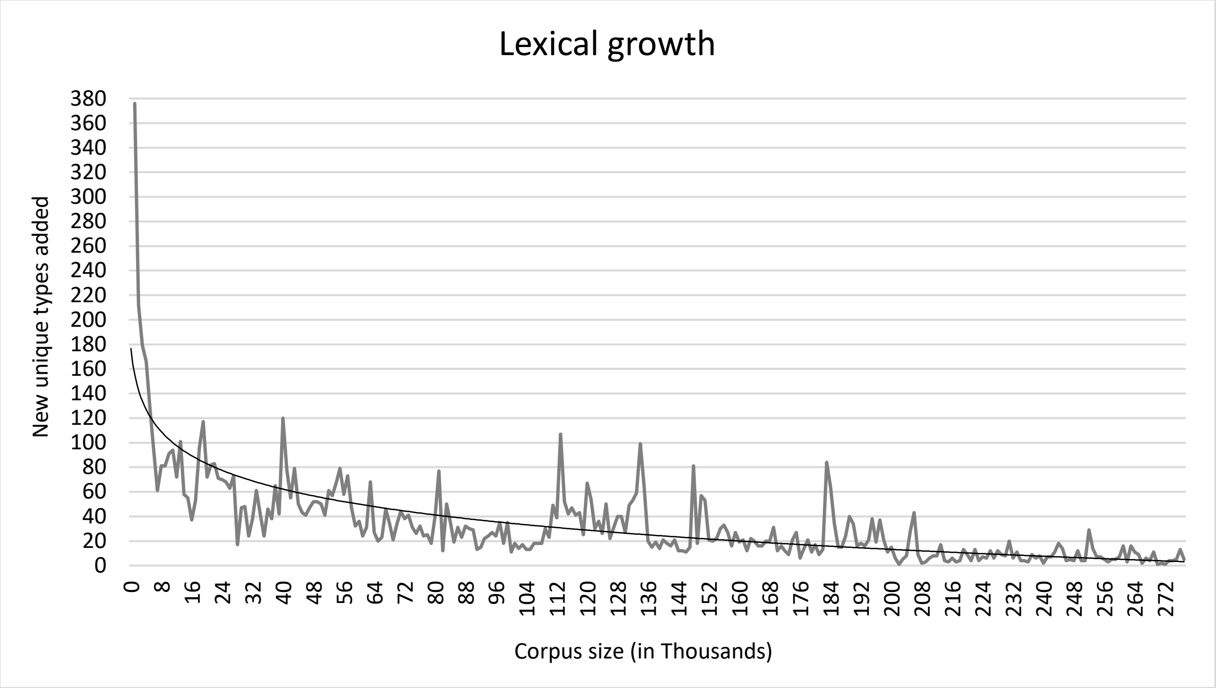
5.3 Lexical closure and corpus representativeness

After having discussed the main terminology of frequency data, the present section will provide a reliability measure for the data contained within the corpus. The use of lexical closure as a measure for corpus representativeness has been discussed and justified in the methodology chapter of this thesis.

Lexical closure is measured by calculating and analysing the growth in the number of types in a corpus, as segments of 1,000 tokens are gradually added to the corpus. Closure is indicated where the number of types “tapers off” gradually as more segments are added (Temnikova et al., 2014).Thus, a corpus can be shown to achieve finiteness, or closure, lexically speaking.

5.3.1 Lexical growth

The following section illustrates the lexical growth within the MDSTB corpus with the purpose of measuring representativeness of the corpus by means of lexical closure. In order to measure the lexical growth within the corpus, all corpus files were compiled into one single corpus text, which was then split into 1,000-word text segments, using the software tool Ant File Splitter (Antony, 2017). Gradually, all segments were added one-by-one into AntConc (Anthony, 2004) to calculate the number of types after every new 1,000 word-segment was added to the corpus. The number of types at each added segment were recorded in an Excel spreadsheet, allowing the calculation of the number of unique types added to the whole corpus at every 1,000 tokens added, thus measuring the lexical growth of the corpus.



Graph 1: Lexical growth of the MDSTB corpus

As can be seen in Graph 1, the number of unique types added to the corpus reduces rapidly with each segment added to the corpus, with the number of new types levelling off for the first time after the corpus reaches a size of 8,000 tokens, where the number of new types added is 81 (8.1%).

Overall, the number of new unique tokens added ranges from 1 to 376, with a median of 33.06 of new unique tokens added per segment overall (with a standard deviation of 36.41). After the corpus reaches a size of 8,000 tokens, this reduces drastically to a maximum of 120 new unique tokens added per segment, with a median of 29.41 (standard deviation of 24.46).

These peaks that indicate an increase in the number of new types added to the corpus. Such spikes tend to correlate with the changes in corpus text, as indicated in Graph 1 above. This is perhaps unsurprising, as each corpus text covers a specific subject matter within Buddhism. Therefore, it is to be expected that every time a new corpus text is added, new lexical items occur within the corpus. Furthermore, stylistic differences between authors and translators may result in a difference in lexical choices made to the style of other corpus texts contained within the corpus. Interestingly, text 7 of the corpus does not appear to contribute much to the lexical growth of the corpus, which indicates that much of the vocabulary of this topic has already been covered by previous corpus texts, supporting further the lexical closure properties of the corpus. Other peaks in the lexical growth occur at corpus sizes of 20,000 tokens, 43,000 tokens and 148,000 tokens, and are not indicative of changes in the texts. Such peaks correlate with changes in the section within a text: in these cases, texts make use of endnote sections, which accounts for the difference in vocabulary introduced here.

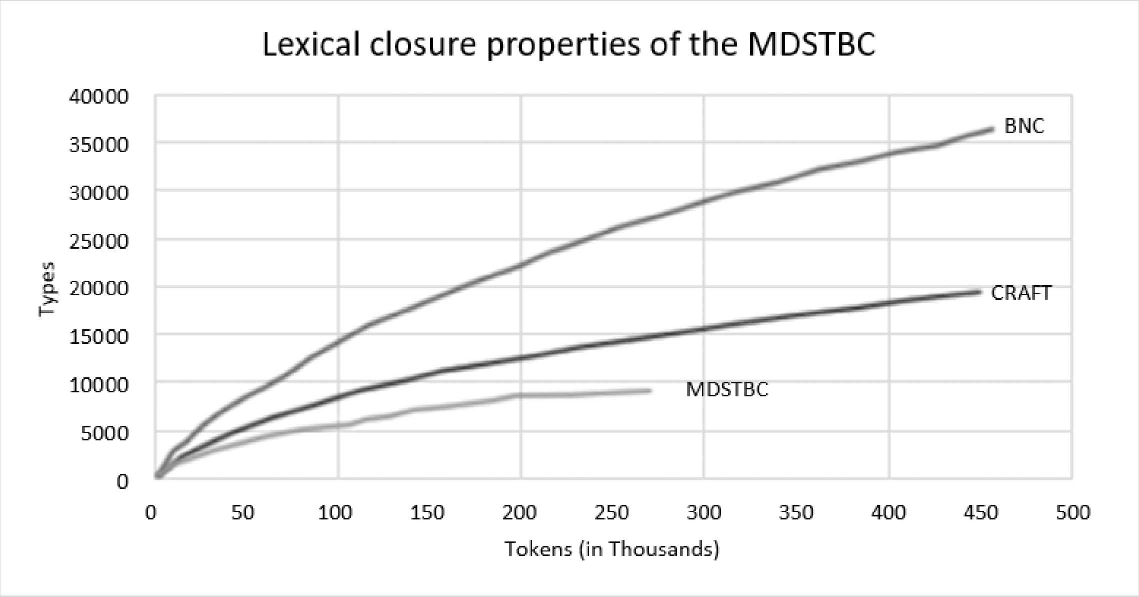
5.3.2 Lexical closure properties of the MDSTB corpus compared to other registers

Overall, it becomes apparent from the data presented above that the written subregister of Tibetan Buddhist *shastra*, as represented in the corpus, shows a tendency towards “finiteness”, thus indicating closure as *per definitionem* of a “specialized domain”:

> Closure is the tendency toward finiteness in a genre or sample of language. It is exemplified, for instance, by limited vocabularies in a specialized domain. If unrestricted natural language tends toward the infinite, then we see the opposite in language samples from restricted domains […]
>
> General language samples will tend to show continued growth in the number of types as long as new tokens are observed – a lack of closure. Sublanguages will show a tapering off in the growth of the number of types after some number of tokens have been observed - in other words, closure (Temnikova, Baumgartner et al., 2014b, p. 1).

In order to understand better the closure properties of the MDSTB, its lexical closure properties have been compared to the findings of the studies conducted by (McEnery & Wilson, 2001) on the BNC as a representative of a general language corpus and by Temnikova et al. (2014) who are also methodologically based on the work by McEnery and Wilson (2001), and who investigated the closure properties of language used in the specialised domain of scientific journal articles from the discipline of molecular biology through the Colorado Richly Annotated Full Text (CRAFT) and GENIA corpora. The CRAFT corpus was found to display similar, yet slightly stronger closure properties to the GENIA corpus (Temnikova et al., 2014), which is why it was chosen as a reference corpus representing a specialised domain here. The BNC was chosen due to the previous work on closure properties conducted on it, and because the written subcorpus thereof will be considered in further analyses within this thesis. Analyses to test the closure properties are not methodologically limited to lexical closure. In addition to lexical closure analysis, other specialised corpora are analysed for their closure properties in terms of parts of speech and sentence type. The present investigation will be limited to lexical closure as the MDSTB corpus has not been annotated for parts of speech, making POS and sentence type closure analyses unfeasible.

Aligned with the methodology proposed by McEnery and Wilson (2001), and subsequently adopted by Termikova et al. (2014), lexical closure properties will be displayed through the type token ratio per thousand tokens added to the corpus.



Graph 2: Lexical Closure properties of the MDSTB corpus compared to the BNC and CRAFT[27]

As can be seen in Graph 2, the MDSTB corpus displays a significantly greater limitation of its vocabulary – its tendency towards “finiteness”, even when compared to the CRAFT corpus which, similar to the MDSTB corpus, is also comprised of one

[27] Closure data from the CRAFT and BNC corpus are based on the study by Temnikova et al. (2014)

single textual subregister (journal article) of a specialised domain (molecular biology). The main caveat is, however, that the textual registers differ from one another. To enhance comparability, it would be interesting to conduct a study in the future considering the same textual register (journal article) of the specialised domain, Buddhism. Unsurprisingly, the BNC, as a general reference corpus, displays no signs of finiteness in its lexical properties, as previously identified.

Evaluating the representativeness of the corpus by means of measuring lexical closure has helped the researcher of the present study generate confidence that the data obtained within the corpus is representative of the language represented therein, and that the addition of a large number of further texts would not significantly alter the results. It thus indicates that the insights gained from the corpus, at least from a lexical perspective, are somewhat typical of the written subregister under investigation. Such an approach to measuring representativeness is, however, limited in that it does not provide information on the number of frequencies for each token but merely on the number of types (i.e. unique tokens) in the corpus. Useful analyses, even of a qualitative nature, that are corpus driven, would require a number of occurrences for hypotheses on or meanings or functionality of language use to be teased out. A low number of observed frequencies of lexical items may yet impede representativeness in terms of a lack of "useful observed frequencies" of vocabularies.

5.4 Lexical repetition

Type-token ratio is used to measure lexical repetition of a corpus, and is calculated by dividing the number of types by the number of tokens of a corpus, with the result expressed as a percentage (Baker et al., 2006, p. 162). Such a measure can indicate the degree of repetition within a corpus, and thus "a high type/token ratio suggests that a text is lexically diverse, whereas a low type/token ratio suggests that there is a lot of repetition of lexical items" within a corpus. This also causes larger corpora to show a tendency towards a lower type/token ratio, "due to the repetitive nature of function words" (Baker et al., 2006, p. 162).

Following this calculation, the TTR of the MDSTB corpus is 3.34%. Interestingly, as can be seen in the Figure 4 below, there is great variance in the lexical repetition within the corpus. This is caused by the variance in text length as the corpus is comprised of full texts. Yet, even at individual text level, the type-token ratio can be considered low for written registers, where the type-token ratio varies between 4.37% and 15.62%.

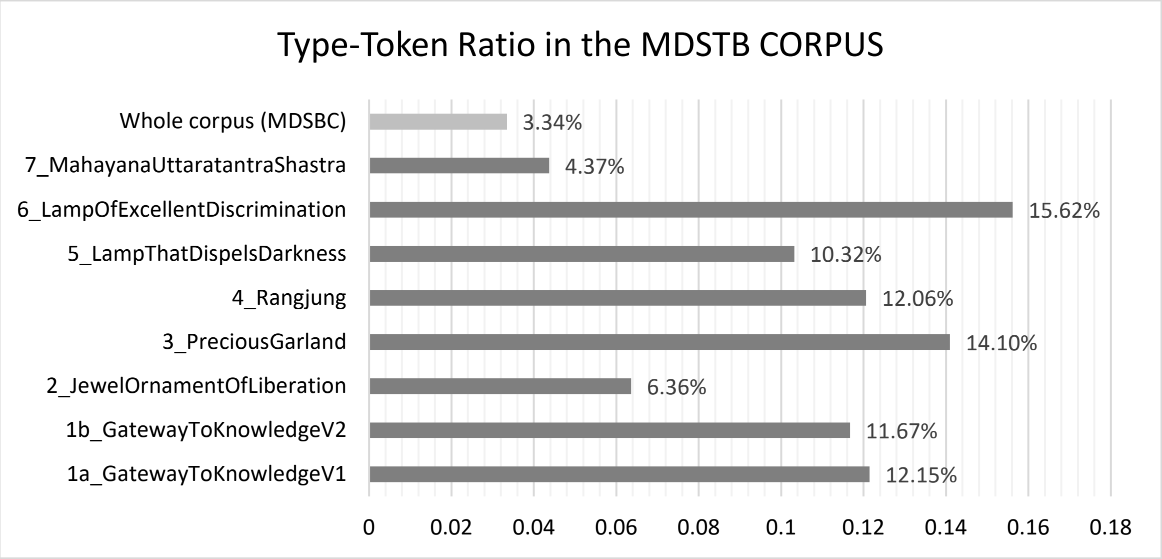


Figure 4: Type-Token Ratio in the MDSTB corpus

This is a surprisingly low ratio for a small specialised corpus of written texts when compared to, for example the BAWE corpus which is 23 times larger in size, and yet has a larger type-token ratio, as can be seen in Table 5 below. One of the reasons for this result is the fact that the BAWE is not limited to one academic discipline, and as such it can be expected that it will include a more diverse vocabulary.

Corpus	**Type-Token Ratio**	**Corpus size (tokens)**
BAWE (written)	5.98	6,506,995
MDSTB (written)	3.34	281,290

Table 5: Lexical repetition of MDSTB corpus compared to BAWE corpus

One reason for the low type-token ratio can be linked back to the lexical closure properties. The earlier a corpus reaches closure, the lower the number of new types in the corpus for each new text added to the corpus, while the number of tokens rises steadily. Such explanation would, however, not justify the significantly lower type-token ratio of the MDSTB corpus when compared to the CRAFT corpus, as has been plotted in the calculations of lexical closure properties in the above section. Here, the MDSTB corpus displays a lower TTR when compared to another register that is comprised of one register of a specialised domain (i.e. journal articles in molecular biology).

One may thus hypothesise that the language used in the *shastras* is indicative of a lower level of formality, solely based on its lexical repetition:

> Production from registers that can be considered to be less formal, such as spoken production, tend to have a lower TTR and more formal registers, such as written production, tend to have a higher TTR (Kaatari & Larsson, 2019, p. 11).

The following section will illustrate that this is not the case for the *shastra* as there is a significant overlap in the distribution of the most frequent items within the wordlist when compared with other written registers, and academic written subregisters in particular (see 5.5 Word frequency lists). As such, the level of formality in this context should not be affected by its low TTR.

5.4.1 Theoretical implications

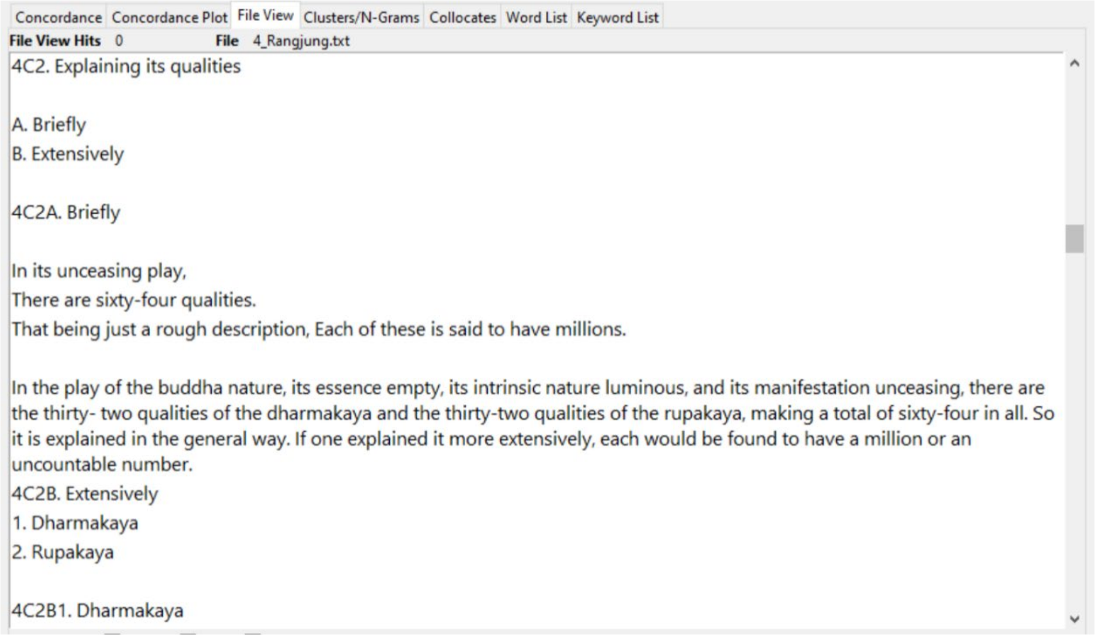
Such analysis clearly indicates the value of the framework for register analysis by Biber and Conrad (2013) with its emphasis on the synthesis of the situational and linguistic analysis of a typical linguistic feature. In the case of the type-token ratio, such an approach has been able to draw the link between the Buddhist practice of the memorisation of *shastras*, and thus provided the rationale behind the feature of TTR and its centrality within the register.

Furthermore, such findings cast a critical light on the patterns and their associations within Dimension 1 of Biber's MD framework. Here he associates high type-token ratios with informative texts, whereas low type-token ratios have been claimed to be indicative of involved production (Biber, 1988). The analysis of type-token ratios in the present study, however, is inconsistent to such classifications. It has been shown that the reason behind the low type-token ratio within the written subregister of shastra cannot be related to the characteristic of involved production but rather is based on the frequent repetition of, for examples, headings and subheadings. Such features themselves have been found to be indicative of other academic written subregisters (Kearsey & Turner, 1999, p. 5), yet the unusually high frequency and high level of hierarchy in the use of such headings and subheadings is causing such high lexical repetition and thus low type-token ratios. As such, it is suggested that the dimensions within the framework may require an extension or review based on the language used in this specific religious context.

Frequency data alone, such as TTR, is at this point limited in its ability to cast light onto the reasons behind such findings, and further qualitative investigation is required. Follow-up investigations into the different corpus texts at full-text level (file view in AntConc (Anthony, 2004) indicated that the reason for the high level of repetition is the structural organisation of the text, with its frequent use of headings, subheadings and overviews of content. This accounts for the highly repetitive vocabulary, and thus the low type-token-ratio of the *shastras*.

5.4.2 Structural organisation of texts

The frequent repetition of headings, subheadings, the use of in-text table-of-contents-style content overviews is illustrated in the screen shots of file views in Figures 5-8. Additionally, such examples indicate the use of the different numeral systems to indicate the hierarchical relationship between different sections in the text. As can be seen, there is some inconsistency in the way these numeral systems are combined to indicate the structure of each text, yet the rigorous structuring of texts appears to be a salient feature of the written subregister of *shastra*.



4C2. Explaining its qualities

A. Briefly
B. Extensively

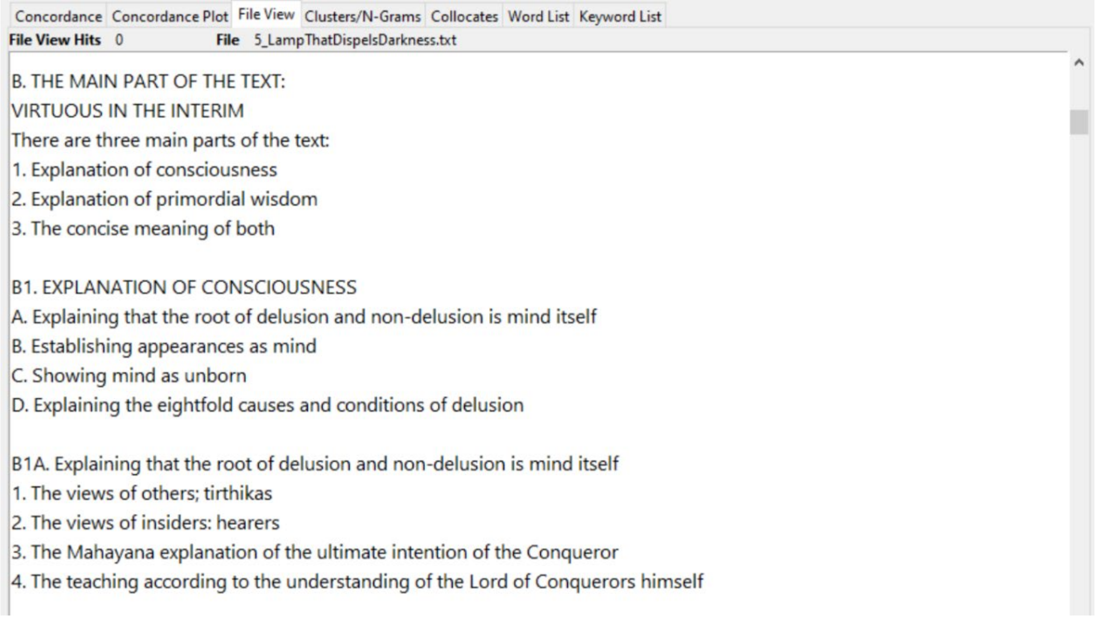
4C2A. Briefly

In its unceasing play,
There are sixty-four qualities.
That being just a rough description, Each of these is said to have millions.

In the play of the buddha nature, its essence empty, its intrinsic nature luminous, and its manifestation unceasing, there are the thirty- two qualities of the dharmakaya and the thirty-two qualities of the rupakaya, making a total of sixty-four in all. So it is explained in the general way. If one explained it more extensively, each would be found to have a million or an uncountable number.
4C2B. Extensively
1. Dharmakaya
2. Rupakaya

4C2B1. Dharmakaya

Figure 5: File View: Mixed decimal and alphabetic numeral system in corpus text 4 of the MDSTB corpus

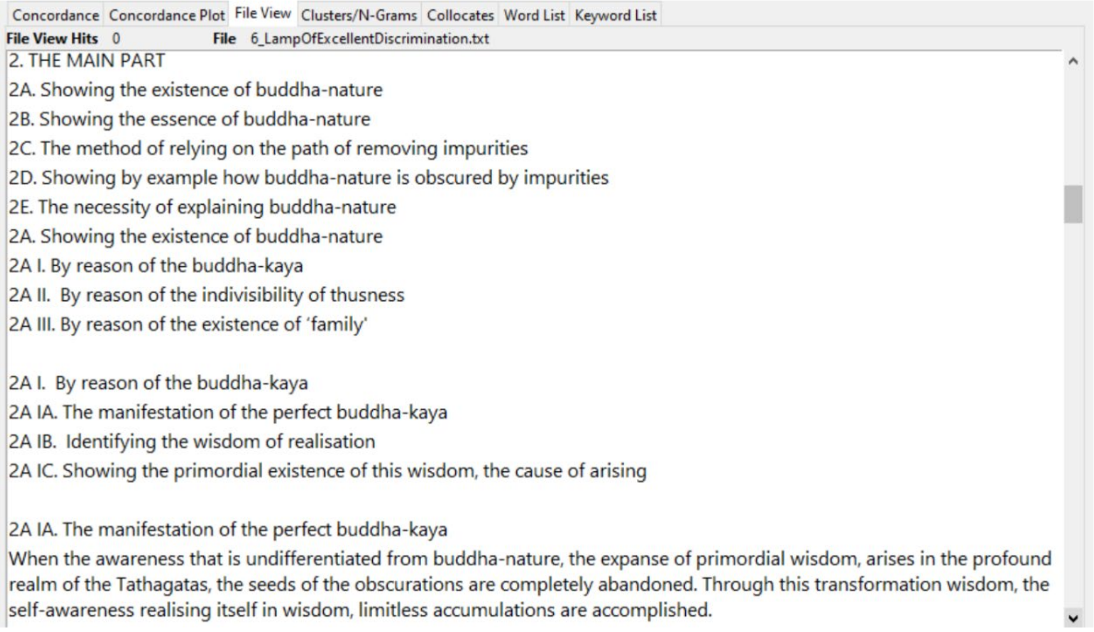


B. THE MAIN PART OF THE TEXT:
VIRTUOUS IN THE INTERIM
There are three main parts of the text:
1. Explanation of consciousness
2. Explanation of primordial wisdom
3. The concise meaning of both

B1. EXPLANATION OF CONSCIOUSNESS
A. Explaining that the root of delusion and non-delusion is mind itself
B. Establishing appearances as mind
C. Showing mind as unborn
D. Explaining the eightfold causes and conditions of delusion

B1A. Explaining that the root of delusion and non-delusion is mind itself
1. The views of others; tirthikas
2. The views of insiders: hearers
3. The Mahayana explanation of the ultimate intention of the Conqueror
4. The teaching according to the understanding of the Lord of Conquerors himself

Figure 6: File View: Mixed decimal and alphabetic numeral system in corpus text 5 of the MDSTB corpus



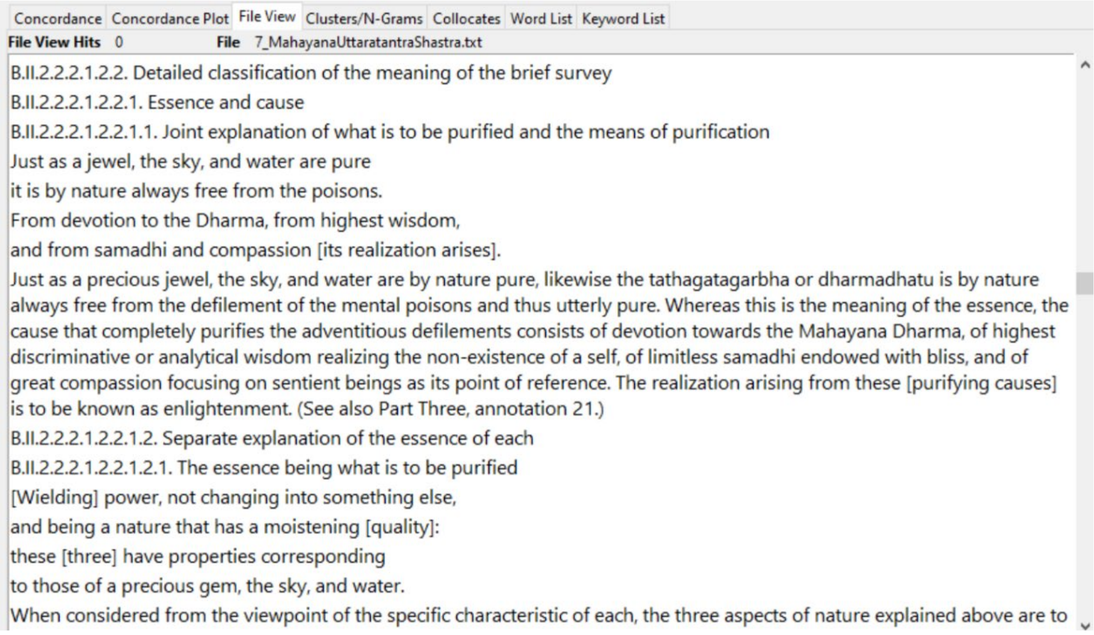
File View Hits 0 File 6_LampOfExcellentDiscrimination.txt

2. THE MAIN PART
2A. Showing the existence of buddha-nature
2B. Showing the essence of buddha-nature
2C. The method of relying on the path of removing impurities
2D. Showing by example how buddha-nature is obscured by impurities
2E. The necessity of explaining buddha-nature
2A. Showing the existence of buddha-nature
2A I. By reason of the buddha-kaya
2A II. By reason of the indivisibility of thusness
2A III. By reason of the existence of 'family'

2A I. By reason of the buddha-kaya
2A IA. The manifestation of the perfect buddha-kaya
2A IB. Identifying the wisdom of realisation
2A IC. Showing the primordial existence of this wisdom, the cause of arising

2A IA. The manifestation of the perfect buddha-kaya
When the awareness that is undifferentiated from buddha-nature, the expanse of primordial wisdom, arises in the profound realm of the Tathagatas, the seeds of the obscurations are completely abandoned. Through this transformation wisdom, the self-awareness realising itself in wisdom, limitless accumulations are accomplished.

Figure 7: File View: Mixed decimal, alphabetic and roman numeral system in corpus text 6 of the MDSTB corpus



File View Hits 0 File 7_MahayanaUttaratantraShastra.txt

B.II.2.2.2.1.2.2. Detailed classification of the meaning of the brief survey
B.II.2.2.2.1.2.2.1. Essence and cause
B.II.2.2.2.1.2.2.1.1. Joint explanation of what is to be purified and the means of purification
Just as a jewel, the sky, and water are pure
it is by nature always free from the poisons.
From devotion to the Dharma, from highest wisdom,
and from samadhi and compassion [its realization arises].
Just as a precious jewel, the sky, and water are by nature pure, likewise the tathagatagarbha or dharmadhatu is by nature always free from the defilement of the mental poisons and thus utterly pure. Whereas this is the meaning of the essence, the cause that completely purifies the adventitious defilements consists of devotion towards the Mahayana Dharma, of highest discriminative or analytical wisdom realizing the non-existence of a self, of limitless samadhi endowed with bliss, and of great compassion focusing on sentient beings as its point of reference. The realization arising from these [purifying causes] is to be known as enlightenment. (See also Part Three, annotation 21.)
B.II.2.2.2.1.2.2.1.2. Separate explanation of the essence of each
B.II.2.2.2.1.2.2.1.2.1. The essence being what is to be purified
[Wielding] power, not changing into something else,
and being a nature that has a moistening [quality]:
these [three] have properties corresponding
to those of a precious gem, the sky, and water.
When considered from the viewpoint of the specific characteristic of each, the three aspects of nature explained above are to

Figure 8: File view: Mixed alphabetic, roman and decimal numeral system in corpus text 7 of the MDSTB corpus

Texts 4 and 5, for example, use the alphabetic numerals to indicate the chapter number, and then indicates hierarchically lower subsections by alternating between the decimal and the alphabetic numerals, for example *B1A*. The level of hierarchy in both texts is up to 5 subsections (text 5) and 7 (text 4) in a chapter. Each chapter and subsection is clearly indicated not only by the use of headings and subheadings, but also by frequent repetition of chapter and subsection level table-of-contents within the main text.

Similarly, in text 6 of the corpus, each chapter and subsection are clearly indicated through the use of headings and subheadings, and through repetition of chapter and subsection level table-of-contents within the main text. The level of hierarchy in this text is up to 5 subsections per chapter. Although the use of the numeral system is similar to corpus text 5, text 6 uses a system that combines decimal and alphabetic numerals, as well as a roman numeral system. Chapters are, in the same way as in texts 4 and 5, indicated by a chapter number, and hierarchically lower subsections use the alphabetic numerals, roman numerals, and then again alphabetic numerals and so on, for example *2A IB*.

Text 7 bears most similarity in its numerical system to text 6. Chapters are indicated by alphabetical numerals; the first subsection is indicated by a roman numeral and remaining hierarchically lower subsections use decimal numerals. Chapter and subsections make use of headings, which are repeated through chapter and subsection level table of contents within the main text. One interesting finding for text 7 is the number of hierarchical levels within a chapter, frequently exceeding 10 subsections.

Numeral systems, headings and subheadings are features to enhance text navigation (through, for example, cross-referencing) and aid the processing of information within written texts. They are a common feature of academic written registers, as indicated in the research by Kearsey and Turner who investigated the use of such structuring devices within science textbooks as a subregister of academic writing (Kearsey & Turner, 1999, p. 5).

It has been shown that *shastras* are commonly[28] highly structured through their systematic use of numeral systems, headings and table of contents-style overviews in text. It has further been shown that there are some discrepancies in the way this is executed in that some texts make use of a decimal numeral system others make use

[28] It has been shown that there is variance within the corpus and that two texts (1a and 1b) in the MDSTB CORPUS differ from the rest of the corpus in their use of numeral systems

of a combination of decimal and alphabetic numerals, and others combine decimal, alphabetic and roman numerals in their system. There is further variance in the levels of hierarchy within each chapter, and it is perhaps interesting to see a religious text that is structured in a way to include up to 10 levels of hierarchy in a chapter.

This frequent repetition of headings and subheadings inevitably leads to lexical repetition within the corpus. In light of the above findings, it is perhaps unsurprising that the type-token ratio of the within the written subregister of Buddhist *shastras* is so unusually low for a corpus of a written language, indicating high levels of lexical repetition.

Furthermore, as has been seen in the file views in the figures above, headings clearly indicate the move structure of the argument communicated within each text, and the mere view of a complete table of contents of a text can provide a good overview of the way the argument develops throughout. This is a point that will be returned to in the following section where, aligned with the framework of register analysis by Biber and Conrad (2013), a link between the linguistic feature of low type-token ratio and the situational context, in this case the Tibetan Buddhist practice of memorisation and debate, will be established.

5.4.3 The practice of memorisation and debate in Buddhism

The practice of memorisation is deeply rooted in the oral transmission of Buddhist texts at the very outset, where the discourses[29] were not written down. Memorisation was essential to enable the transmission of the sutras (the discourses) and their commentaries.

> Monks and nuns would work together to keep these discourses in memory, orally passing them on to the next generation. […] Even after these discourses were committed to paper, memorization remained a standard practice in the monasteries and nunneries of India. (Roiter, 2015, para. 4)

In fact, the different Buddhist lineages, as illustrated in the introduction to this thesis, are a result of the oral transmission of texts. Research into the differences of such

[29] “Discourses” in the context of Buddhism refers to the teachings of the Buddha

oral transmissions has indicated that there is a strong reliability in the texts that have been transmitted orally (Anālayo, 2020), and that the "core doctrine" has been largely unaffected by transmissions, which are based solely on memory, "with extreme accuracy for over two thousand years" (Sujato & Brahmali, 2013, p. 50).

Shastras, which are the texts investigated as part of this study, from an essential part of the monastic curriculum and as such are the types of texts that would, even in the present day, be commonly memorised by monks. It is perhaps surprising that, even though such texts have now been recorded, they are still memorised, and that memorisation still forms an essential part of monastic life (Dreyfus, 2003) as the transmission of Buddhist texts, which are now available in written form, to an ever-increasing extent digitally (Bingenheimer, 2020).

> Memorization is a significant part of a monk's daily schedule, and mainly serves three purposes: memorizing philosophical texts for debate, memorizing prayers and rituals, and memorizing practical, advice-oriented texts (Roiter, 2015, para. 6).

It is further elaborated, on the example schedule of a present day monastery in India, that monks will engage in memorisation of texts in the early morning, in the evenings as well as completing "memorisation retreats", which will engage the monks in memorisation all day for the duration of their holidays (Roiter, 2015). This account is aligned with the insights provided into monastic life by Dreyfus (2003).

There will be variance in terms of the content that is memorised by the individual and may range from single passages to full texts. Such memorisation forms an essential part of preparing and enabling monks to participate in debates, which form another core part of monastic daily life:

> The best memorizers will memorize entire texts […]. Others will focus only on the definitions, divisions and key passages. These texts are often terse and confusing, and will only gradually unfold their meaning through prolonged reflection and debate. At debate, which can last up to six hours a day, one cannot carry a book, so all debating must be done from memory. Often a monk will initiate a public debate with a quote from a text, prompting the defender to correctly identify the source and context. If he cannot answer, the crowd will yell "Chay!" ("Speak!") until he can. If he is stuck, the questioner might give a few more words from the quote. Later on in

> the debate, the questioner is free to give quotes to support his argument,but is expected to recite them from memory (Roiter, 2015, paras. 9-10).

Thus, the practice of debate, heavily reliant on the memorisation of texts, becomes a tool to aid understanding of such texts. The seminal book “The sound of two hands clapping”, which provided first insights into the life of a Buddhist monk to a Western audience by Dreyfus (2003), emphasises the centrality of debate in monastic life as the author chose to use the clap of the hands, which is used in debate by the instigator of the debate, as the title of his book.

Debate, in the context of Buddhism, differs from Western debates. The main aim of Buddhist debate “is to extinguish ignorance and cultivate wisdom” (Wangkhang, 2019, para. 3). Whilst the aim of Western debates is to win the debate, Buddhist debate is “not only about winning but developing compassion by paving the path for your opponent to come to the same conclusion as you” (Wangkhang, 2019, para. 3). The underlying rationale behind the practice of Buddhist debate is analysing and interrogating one’s understanding of a Buddhist subject. Buddhist debate is underpinned by and follows the Buddhist practice of hearing and thinking, i.e. listening to the teaching and critically reflecting upon them before putting them into practice.

> The point of debating is to help one analyze [... . F]irst you listen to subject matter, then you go into a debate court and debate on the points you’ve learned, then you go back to your room and meditate on all of it (Wangkhang, 2019, para. 4).

This critical engagement with texts allows practitioners to “interrogat[e] spiritual practice through a logical lens” (Wangkhang, 2019, para. 25).

The account of the Tibetan-Canadian journalist Rignam Wangkhang who reflected on his engagement in a Buddhist debate (Wangkhang, 2019), provides an interesting insight into the practice. He recalls that significant time is spent in preparation for the debate. After being given a topic for the debate, studying said topic, often in study groups, takes place. This can stretch over a period of months. The focus is on gaining in-depth knowledge and understanding of the key Buddhist teachings on the subject matter, as well as argumentative structures to address different potential responses within the debate:

> We [...] practiced the structure and techniques of debate for months with genla [(teacher)] and the other enthusiastic geshes [(monks who have dedicated their lives to the study of Buddhism)] at Gajang [Buddhist Centre]. We mapped out every potential response. And in the process, we learned Tibetan phrases and Buddhist concepts that we had never heard of before, while earning the respect of our genlas (Wangkhang, 2019, para. 10).

As has been shown above, the memorisation of texts, as a key part of the practice of debate, is a central aspect of Buddhist monastic practice.

5.4.4 Linking type-token ratio and the structural organisation of texts to the Buddhist practice of memorisation

The ability to memorise Buddhist texts, and the argument structure therein, relies on a number of mnemonic devices employed. An overview of devices that were commonly used as part of the early oral transmission of texts have been provided by Sujato and Brahmali (2013, pp. 50-54), and include:

- Repetitions of words, phrases, passages and whole Suttas;
- Concatenation of textual units;
- Formal structures;
- “Summary” and “exposition”, which is a standard feature of Indian oral education;
- Numbered lists

Such mnemonic devices rely heavily on frequent repetition, structuring (through lists and numbering) and memorising sequences, which relates directly to the findings that were presented earlier in this section, where *shastras*, as represented in the MDSTB corpus, are commonly highly structured and repetitive through their systematic use of headings, subheadings, numeral systems and table-of-contents-style overviews in text. Such use of structuring devices will ease the practice of memorisation of the *shastras* as full texts, or initially the move structure of the argument through memorisation of headings. The numeral system will aid the memorisation technique of concatenation of textual units. The frequent lexical repetition of aspects of the argument structure can be directly linked back to the high observed lexical repletion, measured through type-token-ratio, in the present corpus. As such, it is argued that, considering the way such texts are engaged with in the target community, low type-token ratio, caused by the structuring of the texts within the corpus, and the highly repetitive vocabulary, is a pervasive feature of the written subregister *shastra*.

5.5 Word frequency lists

The use of word frequency lists has been used in this study in a three-fold way:

1. To allow the researcher to identify shared features and differences between the written subregister *shastra* and other written registers
2. To position the written subregister *shastra* among other written registers
3. To identify linguistic features within the register for further investigation, thus taking a data-driven approach

position	**MDSTB corpus**	**frequency per million tokens**	**BNC written**	**frequency per million tokens**	**BAWE**	**frequency per million tokens**
1	the	92193	the	64420	the	72059
2	of	53429	of	31109	of	39681
3	and	33382	and	27002	and	30549
4	is	26674	a	22222	to	28047
5	to	19116	in	19466	in	22444
6	in	17302	to	26062	a	19966
7	are	13015	is	9961	is	16293
8	a	12880	that	9936	that	11613
9	it	11593	was	9368	as	9964
10	that	9204	it	9298	for	8719
11	as	8649	for	8815	be	8508
12	by	8145	with	6821	this	7962
13	this	8013	he	6756	it	7502
14	one	7963	be	6742	are	6256
15	not	7917	on	6569	with	6193
16	all	7594	I	6494	on	5949
17	from	7355	as	5621	by	5938
18	be	6296	by	5528	was	5395
19	for	6179	's	4945	not	4901
20	with	5126	at	4868	from	4604

Table 6: 20 most frequent tokens of the MDSTB corpus, BNC written and BAWE corpus; frequency per million tokens

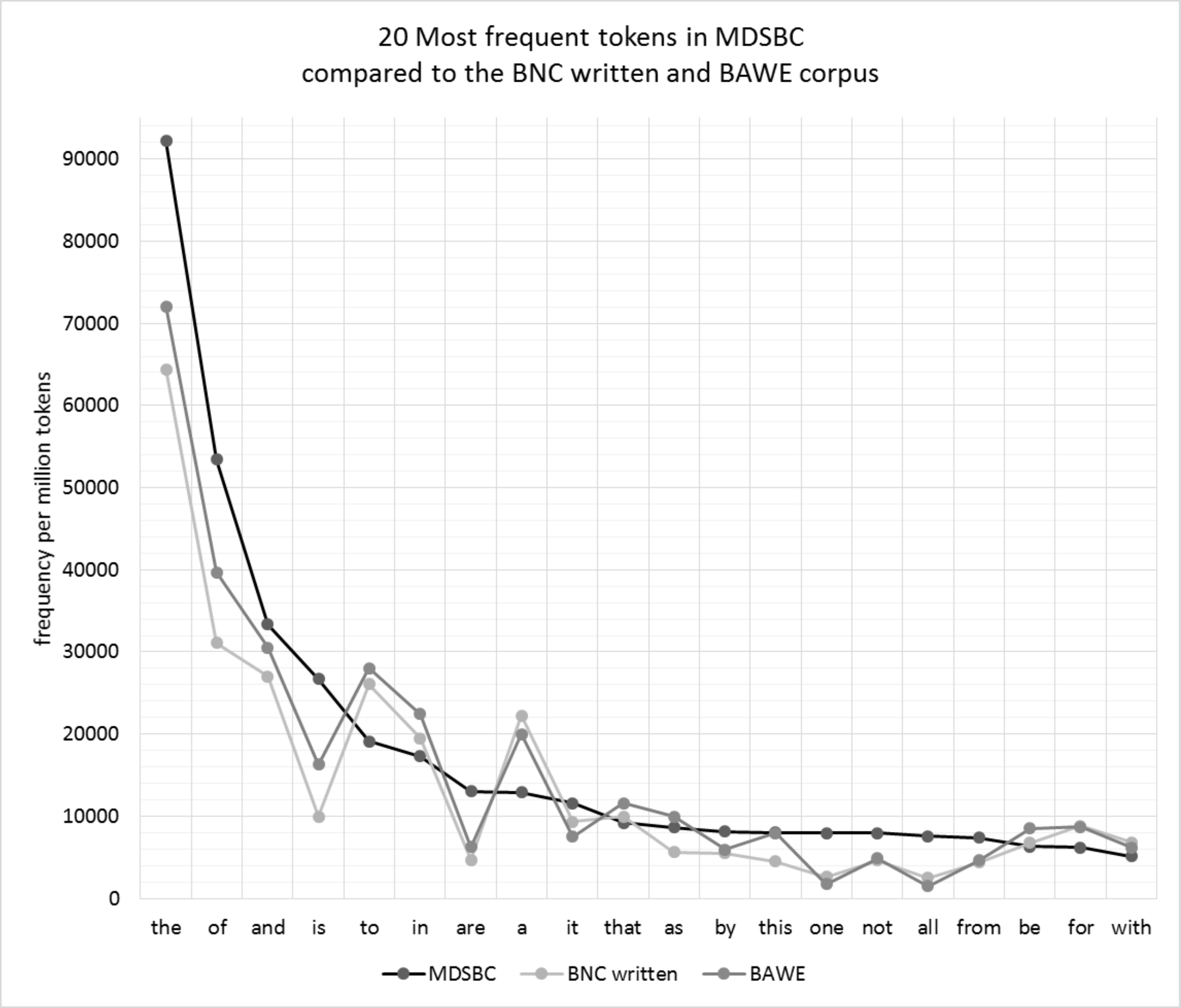
As can be seen from Table 6, the language used in the Mikyo Dorje Shedra Tibetan Buddhist Corpus, in terms of the 20 most frequent tokens, overlaps with general English written language as represented within the BNC written subcorpus in 14 tokens. Most notably is the overlap, unsurprisingly, of the first three most frequent tokens *the, of* and *and.* This overlap is also reflected in comparison with the BAWE.

Furthermore, the most frequent tokens of the MDSTB corpus overlap more strongly with the language used in academic writing as represented in the BAWE corpus, with only two tokens that differ in the top 20 wordlist: *one* and *all*, and *on* and *was*.

Interestingly, *was*, which is highly frequent in both, the BNC (position 9) and the BAWE (position 18), is not highly frequent in the MDSTB corpus (position 155) with a frequency of only 832 tokens per million.

It will be beyond the scope of this thesis to investigate all marked differences in the word frequency list. *One* among the tokens that stands out here, has been selected for further investigation as it is also a positive keykeyword within the corpus, and as such, it is considered a salient feature of the written subregister shastra.

Although the corpora align to a large degree in the distribution and inclusion of tokens within the lists of the 20 most frequent tokens, there are marked differences in the frequency distribution of such tokens, as can be seen in Graph 3:



Graph 3: 20 most frequent tokens in the MDSTB corpus compared to the BNC written and BAWE corpus

	Token	**MDSTB corpus** Frequency per million tokens	**BNC written** Frequency per million tokens	**BAWE** Frequency per million tokens
1	**the**	92193	64420	72059
2	**of**	53429	31109	39681
3	**and**	33382	27002	30549
4	**is**	26674	9961	16293
5	**to**	19116	26062	28047
6	**in**	17302	18978	22444
7	**are**	13015	*4731*	6256
8	**a**	12880	22222	19966
9	**it**	11593	9298	7502
10	**that**	9204	9936	11613
11	**as**	8649	5621	9964
12	**by**	8145	5528	5938
13	**this**	8013	*4506*	7962
14	**one**	7963	2609	1758

15	**not**	7917	*4618*	4901
16	**all**	7594	2486	1517
17	**from**	7355	*4360*	4604
18	**be**	6296	6742	8508
19	**for**	6179	8815	8719
20	**with**	5126	6821	6193

Table 7: 20 most frequent tokens in the MDSTB corpus compared to the BNC written and BAWE corpus; frequency per million tokens

Firstly, it is worth pointing out here that the most frequent tokens of the MDSTB corpus, as indicated in Table 7 above, account for a much larger proportion of the overall corpus than compared to both, the BNC written and the BAWE corpus. This finding again resonates with the low type-token ratio which indicated a high lexical repetition within the corpus, and the unusually high frequency of the top three tokens *the, of* and *and* can be explained by the high frequency in the use of headings, subheadings and lists, that have been alluded to in the previous section, and that will be dealt with in detail in chapter 7.

Secondly, when compared to the general written corpus BNC, *the* is nearly one and a half times as frequent in the MDSTB corpus, *of* is just over one and a half times as frequent and *is* has a frequency of over two and a half times over the frequency in the BNC. High frequency of such items may be indicative of a highly frequent use of prepositional postmodified noun clauses. A concordance search in AntConc using the queries in Table 8 below revealed that this suggested n-gram accounts for 25,348 occurrences per million tokens:

N-gram	**Raw frequency**	**Frequency per million tokens**
the _ of _	7130	25,348
the _ of _ is	339	1,205
the _ of _ are	56	199

Table 8: MDSTB corpus frequency of n-gram "the _ of _ (is/are)"

The frequent use of prepositional postmodified noun phrases such as described above is one feature of written academic English language use. This is confirmed when comparing the frequencies of the MDSTB corpus with the BAWE, where frequencies in the MDSTB corpus for *the* are only 1.26 times higher, *of* are 1.31 times higher, *is* is 1.63 times higher and *are* is just over twice as frequent. This, perhaps, leads to the argument that the language use within the texts is perhaps somewhat more similar to the more formal style of writing of academic written registers than that of written English language use in general. The absence of personal pronouns within the corpus would support this argument. Such initial insights into the language used in *shastras* through the generation of wordlists already allow the hypothesis that *shastras* may be placed along the first dimension of Biber's (1988) Multidimensional Analysis framework, to indicate the informational (rather than involved) characteristic of the register.

It is noteworthy that the frequency of *one* in the MDSTB corpus is over three times higher than in the BNC written subcorpus, and over 4.5 times higher than in the BAWE corpus. Again, such a finding justifies the further investigation into the use and function of *one* in Tibetan Buddhist language use and will be provided in chapter 5.

Further exploration of the word frequency list indicated some unusual findings. From the above analysis, it was expected that the corpus would reveal a relatively low frequency of the personal pronoun *I*, based on the data analysis of the 20 most frequent tokens within the corpus. As can be seen in the extract of Appendix A: Wordlist MDSTBC (200 most frequent items) below, *I* is only in position 51 with a frequency of 2,396 per million tokens.

5.5.1 Wordlist: lexical items

Position	Raw frequency	Frequency per million	Node	Lemmas
1	1254	4458	**buddha**	buddha 1013 buddhas 241
2	1220	4337	**say**	said 366 say 107 saying 21 says 726
3	1192	4238	**beings**	beings 1192
4	1086	3861	**mind**	mind 1010 minded 3 minds 73
5	996	3541	**wisdom**	wisdom 936 wisdoms 60
6	845	3004	**other**	other 454 others 391
7	843	2997	**dharma**	dharma 728 dharmas 115
8	811	2883	**nature**	nature 805 natures 6
9	805	2862	**cause**	cause 477 caused 23 causes 271 causing 34
10	805	2862	**like**	like 801 likes 4
11	741	2634	**quality**	qualities 643 quality 98
12	717	2549	**see**	saw 13 see 251 seeing 212 seen 179 sees 62
13	693	2464	**great**	great 621 greater 64 greatest 8
14	683	2428	**path**	path 591 paths 92
15	624	2218	**mean**	mean 25 means 587 meant 12
16	623	2215	**object**	object 310 objects 313

17	620	2204	**way**	way 528 ways 92
18	606	2154	**sentient**	sentient 606
19	603	2144	**arise**	arise 240 arisen 36 arises 136 arising 182 arose 9
20	552	1962	**bodhisattava**	bodhisattava 1 bodhisattva 302 bodhisattvas 249
21	544	1934	**suffering**	suffering 476 sufferings 68
22	527	1874	**free**	free 462 freed 55 freeing 4 frees 6
23	518	1842	**meaning**	meaning 506 meanings 12
24	500	1778	**body**	bodies 64 body 436
25	483	1717	**practice**	practice 326 practiced 41 practices 54 practicing 62

Table 9: Lemmatised word list of the 25 most frequent lexical items in the MDSTB corpus

The lemmatised wordlist in Table 9 above indicates the overall topical aboutness of the corpus (see also Appendix B: Lemmatised Wordlist MDSTBC (200 most frequent items)). As such, to a general audience, it is perhaps unsurprising that the most frequent node is *BUDDHA*, and other nodes include *BEING, WISDOM, DHARMA, PATH* and *SUFFERING*. Other lexical items may be less expected such as the loanword *BODHISATTVA.* Other items may have been expected to “make” this list, yet are less frequent, such as *MEDITATE* or *NIRVANA*. Although a wordlist of lexical items can provide an interesting first insight into the lexis, and as such give indications to possible semantic fields, it does not account for the distribution of lexical items across the corpus based on topical differences, for example, or author’s style or preference. This is particularly the case for the present corpus, where full texts were sampled, which will inevitably skew the results within a purely frequency based wordlist.

It is for this reason, that a keykeywordlist has been generated. The reference corpus that has been chosen for this is the BNC written subcorpus. Further investigation of distribution of such features across the corpus can also highlight potential pervasive lingusitic features and thus omit those features that may occur more frequently within some corpus texts rather than those that are evenly distributed.

5.5.2 Keykeywordlist

A keywordlist, as defined in the earlier section of this chapter, can provide insights into the “aboutness” of a corpus. By choosing a general written reference corpus

(BNC written), the results will indicate any items within my corpus that are unusually high or low in frequency when compared to the reference corpus. As such, a keywordlist can yield unusually highly frequent grammatical items, and will for the most part indicate lexical items that indicate the aboutness of the corpus.

As is the case for the wordlist generated above, a keyword list will not account for variation in the distributions of word frequencies across the corpus. It is for this reason that the keywordlist was further narrowed down to identify keykeywords (KKW). KKWs are keywords (KWs) within a corpus that are keywords in most or all of the corpus texts contained within the corpus. Thus, KKWs, as opposed to KWs, will not identify tokens that may be overrepresented in few corpus texts but rather be homogenously distributed across the corpus. Thus, KKWs are better placed to help determine the "aboutness" of the corpus as well as register features of the corpus in comparison to the reference corpus. Of course, this keykeywordlist needs to be contextualised with the word list as the keykeywordlist may omit highly frequent items that are also highly frequent in the reference corpus. In this study, not only positive but also negative KWs were considered, and further limited to positive and negative KKWs.

	+KKW	**MDSTB corpus frequency per million tokens**	**Keyness**	**effect size**	**BNC written frequency per million tokens**	**MDSTB corpus compared to BNC frequency**	**BAWE frequency per million tokens**	**MDSTB corpus compared to BAWE frequency**
1	buddha	3573	10590.14	0.0073	0	#N/A	1	4067.94
2	beings	4231	9867.46	0.0085	0	#N/A	34	125.11
3	dharma	2563	8065.77	0.0052	0	#N/A	0	8755.19
4	wisdom	3306	7511.17	0.0067	0	#N/A	19	173.74
5	sentient	2154	6614.89	0.0044	0	#N/A	1	1839.68
6	is	26492	5476.6	0.013	9961	2.66	16293	1.63
7	qualities	2186	4048.52	0.0044	27	80.98	55	39.51
8	mind	3580	3366.21	0.0067	219	16.35	191	18.74
9	the	90554	3165.41	0.0087	64420	1.41	72059	1.26
10	of	52064	3002.8	0.0094	33109	1.57	39681	1.31
11	path	2097	2891.94	0.0042	67	31.31	73	28.60
12	are	12898	2801.5	0.0106	4713	2.74	6256	2.06
13	nature	2848	2782.92	0.0055	195	14.60	496	5.74
14	one	7931	1792.03	0.0086	1839	4.31	2042	3.88
15	all	7586	1696.89	0.0083	2297	3.30	1852	4.1

16	mental	1376	1620.09	0.0027	62	22.19	120	11.45
17	meaning	1482	1527.66	0.0029	72	20.59	260	5.70
18	three	3484	1466.76	0.0057	691	5.04	508	6.86
19	cause	1667	1451.5	0.0032	78	21.38	334	4.99
20	existence	1276	1361.29	0.0025	72	17.73	150	8.48
21	meditation	661	1349.08	0.0013	0	#N/A	6	112.93
22	thus	1859	1278.47	0.0035	228	8.15	658	2.83
23	ultimate	875	1211.51	0.0018	27	32.39	40	22.13
24	attachment	636	1144.8	0.0013	#N/A	#N/A	36	17.60
25	without	2407	1101.42	0.0043	483	4.98	442	5.44
26	object	1099	1072.9	0.0022	57	19.27	139	7.92
27	causes	946	1054.49	0.0019	29	32.61	169	5.6
28	being	3310	1047.74	0.0052	888	3.73	1074	3.08
29	taught	821	942.65	0.0017	39	21.06	35	23.18
30	self	1472	914.31	0.0028	35	42.05	144	10.19
31	arising	629	810.94	0.0013	23	27.36	23	27.56
32	explained	871	761.09	0.0017	76	11.46	109	7.99
33	therefore	1379	711.98	0.0026	241	5.72	1148	1.20
34	truth	814	615.62	0.0016	86	9.47	128	6.35
35	having	1504	615.28	0.0028	341	4.41	311	4.83
36	others	1386	589.28	0.0026	301	4.61	289	4.79
37	through	2332	517.29	0.0038	758	3.08	1031	2.26
38	from	7291	414.78	0.0061	4360	1.67	4604	1.58
39	power	1052	252.97	0.0019	340	3.09	739	1.42
40	neither	398	166.24	0.0008	58	6.86	76	5.27

Table 10: Positive keykeywords; Mikyo Dorje Shedra Tibetan Buddhist Corpus (MDSTB) with BNC written as reference corpus (LogLikelihood); with reference to BAWE frequencies

Overall, 40 types were identified as positive keykeywords, as can be seen in Table 10 above, with an overall raw frequency of 74,568 in the MDSTB corpus, accounting for 26.51% of the tokens of the corpus.

Unsurprisingly, the items computed with the highest statistical significance for keyness are related to the topical “aboutness” of the corpus, thus relating to the semantic field of “Buddhist thought/philosophy”: <*buddha, beings, dharma, wisdom, sentient, qualities, mind, path, nature, mental, cause, existence, mediation, ultimate, attachment, object, causes, being, self, arising, truth, others, power*>*.*

The inclusion of *qualities, cause, causes, ultimate* and *being* may appear unusual in their allocation to this semantic field, but for a member of the target culture the specialised meaning of such tokens within the Buddhist context is predictable. This is the advantage of the analyst, a member of the discourse community, carrying out the analysis of the data, as highlighted in chapter 4. Follow-up research into the context of such tokens will provide further insights. The topical representation of *cause and effect* will be part of the analysis of the use of *one* in chapter 6. *Ultimate, buddha* and *mind* will form part of the analysis of the loanword *bodhicitta* in chapter 7.

5.6 Chapter summary

This chapter has provided an insight into the "aboutness" of the Tibetan Buddhist *shastras* through analysis of its lexical closure properties, lexical repetition and word frequency lists.

Lexical closure in the Mikyo Dorje Shedra of Tibetan Buddhism corpus was calculated and it has been illustrated that, despite the relatively small size of the specialised corpus under investigation in this study, the corpus achieved near closure. It is therefore argued that the corpus is lexically representative of the language it represents. In this way, the present study contributes to the wider discussion of representativeness of corpora, especially of those small in size, and has tested the validity of lexical closure as a reliability measure.

The second part of the chapter measured lexical repetition of the MDSTB corpus by means of type-token ratio (TTR), indicating an unusually low TTR compared to other written registers, particularly given the small size of the MDSTB. This finding was to some degree accounted for by the closure properties of the corpus, but a further investigation to understand fully the low TTR will be required. This example has indicated the limitation of insights that can be gained through quantitative-only investigations and will require a further investigation to uncover fully the reasons behind such high lexical repetition.

The final section of this chapter investigated the language of Tibetan Buddhism through word frequency lists (wordlist, lemmatised lexical wordlist and keykeywordlist), and used such findings as the starting points for further investigations in subsequent chapters. For example, the word frequency list provided the 25 most frequent items within the MDSTB corpus compared to the BAWE and BNC written corpora, and indicated an alignment of the present corpus with other written corpora register, and academic written registers specifically, indicating a formal language use. A further investigation into the use of *one* as an unusually highly frequent token in the MDSTB corpus compared to the BNC and BAWE corpora has been justified by means of a word frequency list as well as a positive keykeywordlist.

CHAPTER 6: *ONE* AND ITS SIGNIFICANCE IN THE EXPRESSION OF BUDDHIST THOUGHT

6.1 Introduction to the chapter

One, a keykeyword in the MDSTB corpus, is a salient feature of Buddhist language use. The present chapter will investigate such significance by means of linguistic and functional analysis, and by recontextualising such findings against the Buddhist philosophical concept of nonexistence of self.

Analysis will be carried out at word level to indicate frequency data, as well as at sentence level (and beyond) by means of concordancing. Findings will be aligned alongside other academic written registers by means of comparison with the Longman Grammar of Spoken and Written English academic subcorpus.

The use of *one* as a substitute pronoun, a generic pronoun and its use as part of proper noun (name) will form the basis of this investigation. Its use as part of a proper noun (name) is frequent in the specific context of Buddhism, and it will be illustrated that such use is indicative of the translational practice of using literal translations of loan words in Buddhist English use, thus posing the challenge of comprehension of Buddhist shastras for an audience outside the target culture.

The use of *one* in its use as a substitute pronoun commonly functions as an anaphoric countable reference and is as such most frequently used in oral conversation and less frequent in written registers. It will be illustrated that the unusually high frequency of *one* in this context not only functions as a cohesive device in general, but more specifically utilises the “countable referencing properties” of this device to cross-reference to concepts that are comprised of components which are numbered. As such, the argument will be made that its use refers to the Buddhist practice of memorisation.[30]

The use of *one* as a generic pronoun will investigate its use as part of different syntactic functions (i.e. sentence object or subject), and its use as the subject within conditional subordination structures. It will be shown that the use of *one* is centred

[30] See 5.4 Lexical repetition

around the topic of “cause and effect” and it will be argued that its use as the subject provides agency to *one* and as such implicitly communicates empowerment to the reader. Further, it will be argued that the use of *one* will indicate objectivity and thus universal applicability of the content communicated and overcome the dualism that can easily be inferred from substituting *one* with personal pronouns. As such, it will be argued that the use of *one* within the corpus becomes a vehicle to reflect the Buddhist thought the written subregister *shastra* aims to communicate.

6.2 Frequency of *one*

As has been identified in chapter 4, when comparing the twenty most frequent items in the corpora, *one* appears in position 14 in the MDSTB corpus, whereas is not represented within the top twenty in either, the BNC written subcorpus or the BAWE corpus. Its frequency in the MDSTB corpus (7963 occurrences per million tokens) is three times higher when compared to the BNC (2609 occurrences per million tokens), and 4.5 times higher than in the BAWE corpus (1758 occurrences per million tokens). Furthermore, *one* is a keyword in the MDSTB corpus when compared to the BNC written subcorpus, and also when compared to the BAWE corpus, and it is a keyword in all corpus texts contained within the MDSTB corpus when compared to the BNC written, thus can be identified as a positive keykeyword (Scott, 1997).

	+KKW	**MDSTB frequency (per million tokens)**	**Keyness**	**effect size**	**BNC written frequency (per million tokens)**	**BAWE frequency (per million tokens)**
14	one	7931	1792.03	0.0086	1839	2042

Table 11: Positive key key word one*; Mikyo Dorje Shedra Tibetan Buddhist Corpus (MDSTB corpus) with BNC written as reference corpus (LogLikelihood); with reference to BAWE frequencies*

The use of *one* in the present corpus, it is argued, is a salient linguistic feature in the written subregister of *shastra* and will thus be further investigated.

6.3 Uses of *one*

One can occupy a number of different uses – as can be seen in the figure and table below. The uses considered as part of the analysis are aligned with the analysis of *one* based the LGSWE corpus (Biber et al., 1999, pp. 351-354) to enable a comparison of the data. Here, *one* has three main identified uses: as a numeral, and as an indefinite pronoun with a substitute or a generic use. Data from the MDSTB will further provide insight into the uses as part of names (proper nouns), which is a frequent feature of the language used within Buddhist context. Further uses have been identified, such as the use of *one* as part of compound adjective, adverbs or nouns. Analysis of such will be beyond the scope of this study and has been omitted for reasons of low frequency in the distribution of uses overall, accounting for only 1.6% of occurrences of *one* across the corpus. The numeral use of *one*, though more frequent within representation of uses (13.5%) will also not be considered within this study as its functionality does not make a valuable contribution to the overall argument of this thesis.

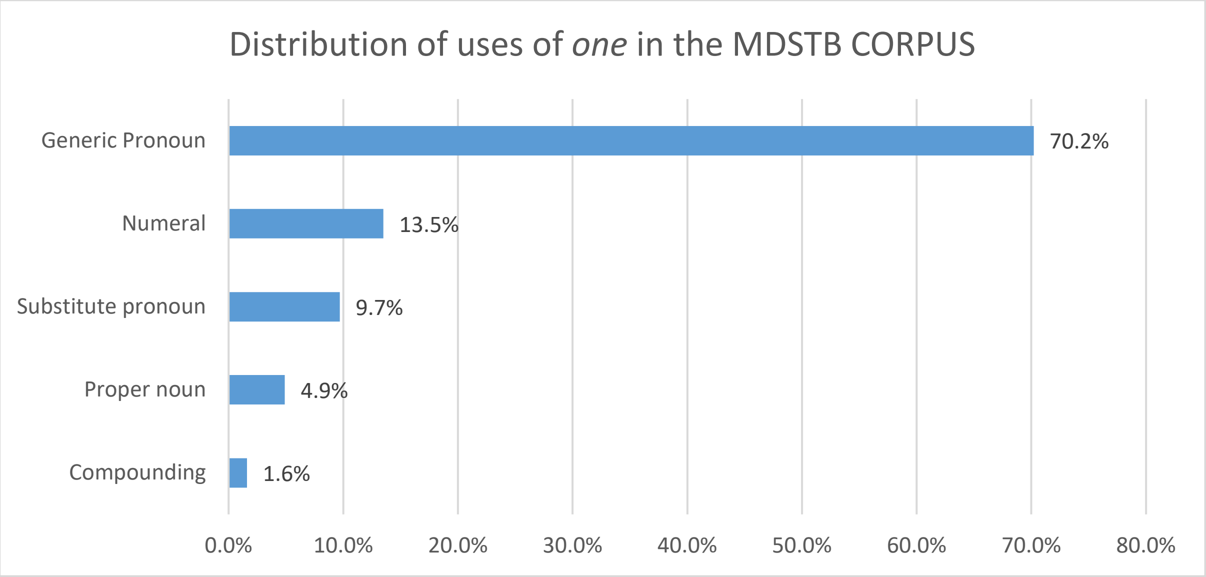


Figure 9: Distribution of uses of one *in the MDSTB corpus*

	frequency per million tokens in (MDSTB corpus)	Frequency per million tokens (LSWE corpus academic)	Distribution of uses (%)
Numeral	**1076**		**13.5%**
Compounding	**117**		**1.6%**
compounded adverb	39		0.5%
compounded adjective	76		1.0%
compounded noun	12		0.1%
Substitute pronoun	**775**	**<125[31]**	**9.7%**
Proper noun (name)	**393**		**4.9%**
Generic Pronoun	**5212**	**>400**	**70.2%**
Possessive pronoun	697		9.4%
Object	252		3.4%
Subject (or Agent)	4259		57.4%
Total	**7963**		**100.0%**

Table 12: Uses of one *in the MDSTB corpus compared to LSWE corpus, and distribution of uses across the corpus*

Use	**#**	**Example from MDSTB corpus**
Generic pronoun	*1*	*The cause through which one accomplishes the four qualities mentioned above is primordial wisdom…*
	2	*First, if one acts with hatred, one will be born in the hell realm.*
	3	*… and the passage of time brings one closer to death*
	4	*Understanding of them will cause one to be learned in the meaning of what is correct or incorrect.*
Substitute pronoun	5	*Meditate on the first one in these ways*
	6	*The both unpurified and purified phase is the one in which the tathagata-garbha is not completely purified from all [defilements]*
	7	*Seven faculties have [physical] form. One is a 'main mind' [primary cognition].*
	8	*These five comprise the training in aspiration bodhicitta. The first one is the method for not losing bodhicitta. The second one is…*

[31] There is no indication of the overall frequency of the substitute *one* in the LSWE but a mere indication of frequency of each form of the substitute *one*, which has been added up as the sum of 4 forms of 25 or less occurrences and 1 form of 50 occurrences. Thus, the actual frequency is likely to be lower than the here indicated 125 instances per million tokens.

Proper nouns (names)	*9*	*the Buddha is the images of the Thus-gone One, the Dharma is the Mahayana scripture, and the Sangha is the community*
	10	*In order to achieve the primordial wisdom of the Omniscient One, you must cultivate great perseverance*
	11	*And even those noble heirs of the Victorious One who dwell on the tenth bodhisattva level are like newborn children*

Table 13: Examples of different uses of one *in the MDSTB corpus*

6.3.1 Proper nouns (names)

This use of *One* as part of proper nouns (name) appears to be common in the specialised context of Buddhist language. *One* here is part of a compound to represent the Buddha. Buddhas are commonly referred to by their Sanskrit names or, at times, through a synonym which, as can be seen in examples 9-11 above, is comprised of three components: *the* + adjective + *One*. Orthographically, this can easily be identified through the use of upper case in the corpus.

All three examples above make reference to the Buddha in their naming, highlighting different aspects of Buddhahood. For example, *the Omniscient One* (examples 10) highlights the characteristic of knowledge and wisdom when referring to the Buddha.

The Thus-gone One from example 9 is a synonym of tathāgata (Pali), which is synonymous for the Buddha.[32] The name "thus-gone" here is a literal translation of tathagata, commonly used in translated texts. The underlying thought is that the Buddha, upon reaching enlightenment, has gone beyond the "endless cycle of rebirth and death" (Chalmers, 1898).

This use of proper nouns (name) provides an excellent insight into the challenges that the Buddhist language can pose with regards to comprehension of such texts by non-specialist audiences (Griffiths, 1981), which has previously been addressed in chapter 1: introduction and chapter 2: literature review.

[32] There are uncountable Buddhas but, when reference is made to "the Buddha" in a non-Buddhist and Buddhist context, reference is made to Shakyamuni Buddha, also called Guatama Buddha, amongst other names.

6.3.2 Substitute pronoun

All substitute pronouns *one* provide anaphoric references. Example 6 indicates such referencing within the sentence boundary, where *one* replaces *phase.* The other examples, 5, 7 and 8 indicate anaphoric referencing beyond the sentence boundary as a means to indicate a count of a concept or phenomenon that links back to an overall item as part of a list:

- referring back to one of multiple statements: *the first one* (example 5);
- referring back 7 faculties: *one is a…* (example 7)
- referring back to 5 virtues: *The first one is the method for not losing bodhicitta. The second one is…* (example 8)

In example 5, *one* replaces *statement*, in example 7 it replaces *faculty*, and in example 8 it replaces *virtue.*

As such, the use of the substitute pronoun *one* contributes to text cohesion by linking concepts beyond sentence boundaries. Within the context of Tibetan Buddhist *shastras*, *one* in its substitute pronoun use is most frequently used (35% of occurrences) as part of listing items or concepts, whereby initial mention is made to the node, which is then followed up by cardinal number and the substitute pronoun *one,* illustrated in example 8 above. A full list of such use can be seen in Appendix C: Concordances of substitute pronoun *one* in the MDSTBC. Furthermore, the feature of substitute *one* is that it “provides a general means of countable reference" that implicitly provides text cohesion and is thus more commonly used in spoken conversation (Biber et al., 1999, p. 334). Interestingly, its frequency in the MDSTB corpus is significantly higher (775 per million tokens) than its frequency in the academic subcorpus of the LGSWE corpus (<125[33] per million tokens).

As such, the use of *one* as a substitute pronoun in the MDSTB corpus not only contributes to text cohesion but is also utilises the “countable referencing properties” of this device to cross-reference to concepts that are comprised of components which are numbered, such as *5 virtues* or *7 faculties*.

[33] There is no indication of the overall frequency of the substitute *one* in the LSWE but a mere indication of frequency of each form of the substitute *one*, which has been added up as the sum of 4 forms of 25 or less occurrences and 1 form of 50 occurrences. Thus, the actual frequency is likely to be lower than the here indicated 125 instances per million tokens.

Such use of structuring and numbering is an important device in the Buddhist practice of memorisation, which requires monks to memorise complete philosophical texts, sometimes verbatim, sometimes focused on the structure of the argument. It has been indicated that lists are an important mnemonic device to aid monks in their practice.[34] As such, this use of *one* aligns with the analysis of chapter 7, where the use of numeral systems as a means to aid memorisation has been dealt with in-depth.

6.3.3 Generic Pronoun *one*

Within written registers, the use of the indefinite pronoun *one* is chosen to replace personal pronoun *we*, *you* and *they*. The use of *one* in this way has been identified as a unique feature of written registers as it enables a more objective form of expression compared to the use of personal pronouns (Biber et al., 1999, p.331). Overall, the use of *one* as a generic pronoun is the most frequent use within the corpus and accounts for 70.2% of occurrences.

Examples 1-4 above illustrate how the use of *one* creates an objective, de-personalised writing style. Using the generic *one* rather than the personal pronouns *I, you, we, us*, which could replace *one* in the examples, would make the texts more interactive and subjective. Such finding is indicative of the level of formality of the written subregister shastra and indicates an alignment along Dimension 1 of the MD analysis towards “information” rather than “interactive” texts.

9.4% of its occurrence (697 per million tokens) is in the possessive form (*one’s)*. In terms of syntactic function, its function as the subject of a sentence (or agent in the passive voice, 4259 per million tokens), such as in examples 1 and 2. Its syntactic function as the subject is 17 times higher than its function as the object of a sentence (252 per million tokens), which is illustrated in examples 2 and 3.

Interestingly, the occurrence of the generic pronoun *one* (5215 per million tokens) is 13 times higher than compared to the Academic written subcorpus of the LGSWE (>400 per million tokens).

[34] An in-depth account of this argument can be found in chapter 7 of this thesis.

6.3.3.1 Internal variation in the distribution of the generic pronoun *one*

This data below suggests internal variation in the way *one* is used in its syntactic function (subject or object of a sentence) within the MDSTB corpus. 48% of all occurrences appear within corpus text 2, predominantly in the syntactic function as subject, as opposed to only 13% being accounted for by corpus text 5, 12% by text 7, 11% by text 1b, 6% by texts 4 and 1a, 2% by text 3, and 1% by corpus text 6. This results in a median of 12.4% per text with a standard deviation of 15.07.

Frequency million tokens	**Object**	**Subject**	**Total per text**
1a_GatewayToKnowledgeV1.txt	50	206	256
1b_GatewayToKnowledgeV2.txt	82	412	494
2_JewelOrnamentOfLiberation.txt	71	2108	2179
3_Gampopa_PreciousGarland.txt	4	96	100
4_Rangjung.txt	4	281	284
5_LampThatDispelsDarkness.txt	14	594	608
6_LampOfExcellentDiscrimination.txt	0	39	39
7_MahayanaUttaratantraShastra.txt	28	523	551
Total	**252**	**4259**	**4511**

Table 14: Distribution of the syntactic functions of the generic pronoun one *in the MDSTB corpus*

Such frequencies, however, provide a somewhat distorted picture as the corpus contains full texts with varying lengths. As such, the frequencies need to be adjusted to indicate a more accurate image of the relative distribution across the corpus, as can be seen in the table below.

ADJUSTED FREQUENCIES (relative distribution)	**Object**	**Subject**	**Total per text**
1a_GatewayToKnowledgeV1.txt	2.03%	8.42%	10.46%
1b_GatewayToKnowledgeV2.txt	2.82%	14.21%	17.03%
2_JewelOrnamentOfLiberation.txt	0.84%	24.87%	25.71%
3_Gampopa_PreciousGarland.txt	0.12%	3.28%	3.40%
4_Rangjung.txt	0.21%	16.72%	16.94%
5_LampThatDispelsDarkness.txt	0.41%	16.99%	17.40%
6_LampOfExcellentDiscrimination.txt	0.00%	4.40%	4.40%
7_MahayanaUttaratantraShastra.txt	0.24%	4.42%	4.66%
Grand Total	**6.67%**	**93.33%**	**100.00%**

Table 15: Relative distribution of the syntactic functions of the generic pronoun one *in the MDSTB corpus; adjusted frequency*

The calculation of the distribution of the general pronoun *one* by means of adjusted frequencies, yields a more even distribution across the corpus, with a range of 2.71% (text 2) to 3.4% (text 3), resulting in an adjusted median of 12.5% with a significantly lower standard deviation of 8.04. Such calculation indicates that *one* is expected to be frequently used across the corpus if the texts within the corpus were of equal length, yet some variation would be expected to persist.

Interestingly, the frequency adjustment has made little difference to the expected use of *one* in terms of its syntactic function. The observed data indicates that in 94.41% of the instances, *one* is used as the subject in a sentence, which has decreased slightly to 93.33% after the data has been adjusted. Accordingly, 5.58% of instances indicated the use of *one* as the object of the sentence in the observed data, which increased to 6.67% post adjustment.

6.3.3.2 Generic pronoun *one* as the object in a sentence

The generic pronoun one in its function as the object of a sentence is extremely infrequent in the MDSTB corpus as compared to its function as a subject, as has been indicated in the section above. Analysis of the 72 occurrences within the corpus indicated that there is a tendency of *one* to be used in the context of expression cause and effect (61 instances), see also Appendix D: Generic pronoun *one* as the sentence object. This is further exemplified in the concordances below, where examples 1, 2, 4, 5 and 6 indicate a causal relationship between an action, mental state or cognitive engagement and its effects: example 5 indicates that by taking a *bodhisattva's vow* (action), the result or effect will be *to benefit others.* Example 6 indicates that by feeling *fury* (mental state), one will *prepare to harm others*. Similarly, example 4 indicates that gaining *understanding* (cognitive engagement) will lead to an understanding of the truth (*what is correct or incorrect*).

There is no tendency within the corpus regarding whether the effect (or impact) on *one* is negative or positive: In 31 instances, *one* was the “recipient” of a positive effect, in 32 instances of a negative effect. Such is also reflected in the examples below. Examples 1 and 2 illustrate a positive effect on *one*, and examples 4 and 5 illustrate a negative effect. Example 3 expresses a cause-effect that has been labelled here as neutral[35] as it expresses a cause-effect relationship where something may be beneficial to others and is unpleasant for *one* at the same time.

1	his life. Further, it nourishes faith, supports perseverance, and quickly frees	one from attachment and hatred. It becomes a cause for the realization of
2	*1 The Ornament of Mahayana Sutra says: Perseverance will liberate	one from the view of the transitory aggregates. If one has perseverance, one
3	finger can be of benefit, Buddha said that even if it makes	one uncomfortable, Helpful things should be done. You should not give traps or
4	schools according to the inner science. Understanding of them will cause	one to be learned in the meaning of what is correct or incorrect. [5,14]
5	harm to others and having harmful motives. The bodhisattva's vow causes	one to benefit others. Without avoiding harm, there is no method of benefiting
6	subsidiary disturbing emotions: [1,75] Fury is the increase of anger. It causes	one to prepare to harm others, such as by hitting them. [1,76] Resentment belongs

Table 16: Concordances of the generic pronoun one *as sentence object*

The semantic category of "cause and effect" is most frequently associated with the use of *one* as the object of a sentence. Given the purpose of Buddhist texts, or in fact any religious texts, in very simplistic terms, to free oneself from suffering (or in the specific context of Tibetan Buddhism where the ultimate goal is to free all beings from suffering) by providing them with a moral or behavioural framework to help them achieve this,[36] the passive role of *one* within the cause and effect relationship is perhaps somewhat surprising and not the most effective way to motivate action. I would argue that, in the way *one* is used as the object of the sentence, a degree of agency is removed from the actor *one* (the person who is intended to adjust their behaviour), and instead implicitly becomes a passive recipient of the effects that either cause benefit or harm. As such, it is perhaps unsurprising that *one* is so infrequently used as the object of the sentence in its generic pronoun use compared to its use as the subject.

[35] One could have made the argument here that the benefit to others outweigh the benefit to *one* but the decision was made not to utilise such ethical or moral considerations into the categorisation process.

[36] See section 5.4 which will discuss the implications on Buddhist thought

6.3.3.3 The generic pronoun *one* as the subject in a sentence

Similarly to the use of *one* as the object of a sentence, the use of *one* as the subject is most commonly[37] in the context of the topic of cause and effect.

Of the 4259 instances (per million tokens) of the use of the generic *one* as the sentence subject, 1959 instances (per million tokens) express cause and effect. As such, it accounts for 46% of all occurrences.

6.3.3.3.1 Conditional subordination

Analysis of concordances of *one* in this context indicated that the most frequent syntactic structure in the corpus to express cause and effect as part of the use of *one* as the generic pronominal, is the conditional subordination (683 occurrences per million tokens; see Appendix E: The generic pronoun *one* as the subject of a sentence in conditional subordination), most commonly with the subordinator *if* (459 occurrences per million tokens) and less frequently with *when* (224 occurrences per million tokens)), to express an inevitable effect that is the result of a specific cause (604 occurrences per million tokens) or to express the definitive outcome of a hypothetical cause (79 occurrences per million tokens). As such, conditional subordination accounts for 23% of all instances where the generic *one* expresses cause and effect.

1	afflicting emotions, by the frequency, and by the object. First, if	one acts with hatred, one will be born in the hell realm. If
2	and touch. From what causes and conditions do these arise? If	one analyses them, one ascertains that the views of worldly people, tirthikas,
3	. Relating to object, one will be born in the hell realm if	one acts nonvirtuously toward beings of higher status; if toward mediocre
4	and establishing the happiness of all sentient beings. When	one attains Buddhahood, there are no conceptual thoughts or efforts. Th

Table 17: Concordances of one *as the subject of conditional subordinate clauses expressing causes and conditions*

As can be seen in the above concordances, *one* is used in such instances as part of the subordinate clause *when one* (see example 4) or *if one* (see examples 1-3). In

[37] The second most frequent use is to give advice (*one should)* or express an obligation (*one has to, one must)* but an investigation thereof is beyond the scope of this thesis, yet would make an interesting topic for future research.

this function, *one* (i.e. the person committing the mental or physical action) becomes part of or controls the cause or condition that leads to either a favourable or unfavourable outcome, or effect. This is particularly the case for examples 1-3 where, through the use of *if,* a hypothetical scenario is presented. In this way, *one* is given agency, and I will argue that such use of *one* has the effect of implicitly communicating empowerment to the reader.

Where the temporal *when* is used, the causes associated with *one* tend to be more of an aspirational nature. This is exemplified through item 4 in the above concordance list, where the *attainment of Buddhahood* is the cause for further effects to arise.

Unlike the use of *one* as the sentence object, the use of *one* as the subject of the sentence indicates a slightly stronger tendency towards positive effect as the result of a cause. The purpose of such expressions of cause and effect is to illustrate the path towards enlightenment for all sentient beings[38], and as such to illustrate causes and effects that will either lead towards the achievement thereof, or causes that will hinder the achievement thereof.

6.3.3.3.2 Causes and effects

Through analysis of concordances, this section will classify the use of *one* in terms of its agency, i.e. whether *one* is associated with the cause of a subsequent effect, whether it is associated with the effect of a previous cause, or whether *one* is preventing an effect from taking place.

As highlighted in the section above, *if* subordination frequently preceeds *one*. This can further be seen in the concordances below. Interestingly, within this structure, *one* is associated with the cause (examples 16-28). Of these, examples 17-22 express a cause resulting in a negative effect, and examples 23-28 express positive effects as part of the cause with which *one* is associated. Notable here are examples 11, which associates *one* with the cause of a negative outcome, yet does so in a simple declarative sentence: *instead, one creates non-virtue which completely binds*

[38] The ultimate goal of Tibetan Buddhism is not the achievement of enlightenment for oneself but for all sentient beings – and as such is not limited to humans.

one to misery. Here, the cause is the generation of an unfavourable condition, which in turn leads to a negative effect.

Causes for negative effects can be summarised as follows:

- Generation of an unfavourable condition (ex. 11)
- Negative mental act (ex. 16)
- Negative mental and physical act (ex. 17, 18, 22)
- Negative physical act (ex. 20, 21)
- Negative motivation (ex. 19)

Causes for positive effects are:

- Positive mental act (ex. 23-27)
- Positive mental and physical act (ex. 28)
- Generating merit (ex. 10)

Examples 12-15 are notable in that they indicate *one* as not being impacted positively as the causes for such effects have not been generated:

- Lack of generating merit (ex. 12, 13)
- Lack of mental achievement (ex. 14, 15)

Other examples associate *one* with effects rather than causes, and positive effects associated with *one* are

- Achieving enlightened states (ex. 1-4)
- Generating merit (ex. 9)

Negative effects associated with *one* are

- Attained negative mental state (ex. 5)
- Attained unfavourable rebirth (ex. 6)

Such categorisation indicates that *one,* which in its use replaces the more interactional personal pronouns *I, you,* and *we*, possesses agency in the achievement of favourable (or unfavourable) effects. This raises the question as to why the generic pronoun *one* is used in favour of the personal pronouns listed above.

Table 18: Concordances of the generic pronoun one *as sentence object*

1	ditative equipoise. Through eliminating the veil of the hindrances to knowledge	one attains the buddhakaya, which has all supreme aspects of qualities. The means	effect
2	the desire realms; and iii) 'nontransferring actions' are the virtues that make	one attain the two upper realms. They are so called since, apart from	effect
3	objects as being existent is the root of samsara, then won’t	one be liberated from samsara if one believes in nonexistence? This latter view	effect
4	cause of merit; one will not fall into the lower realms; and	one quickly achieves the perfect enlightenment. C. Pratimoksa Precepts. The third	effect
5	neficial for oneself. But if one develops loving-kindness and compassion, then	one is attached to sentient beings and dares not attain liberation only for	effect
6	there were evil thoughts By one or two instances of negative speech,	one experiences suffering for 500 lifetimes, and so forth The Verses Spoken Intenti	effect
7	Letter to a Friend says: Even if one became a universal monarch,	One would fall into slavery in samsara. Not only that, even one who	effect
8	– is generated. Vasubandhu says: From birth come actions and defilements, And	one goes on to the next life. The cycle of becoming is beginningless.	effect
9	will obtain limitless merits, f) all the Buddhas will be pleased, g)	one becomes useful to all sentient beings, and h) one quickly attains perfect	effect
10	, a great result will ripen. The Verses Spoken Intentionally say: Even if	one creates small merit, It will lead to great happiness in the next	cause
11	for the three realms and so does not engage in virtue; instead,	one creates nonvirtue which completely binds one to misery [in subsequent lives].	cause
12	, One cannot achieve clairvoyance. Likewise, without the power of clairvoyance,	One cannot benefit sentient beings. Again, without meditative concentration you can	non-effect
13	being What kind of rebirth shall I take? Second, with these faults	one cannot benefit others. Thus, it is said: For should it ever happen,	non-effect
14	for the Path to Enlightenment says: Without the accomplishment of calm abiding,	One cannot achieve clairvoyance. Likewise, without the power of clairvoyance, One c	non-effect
15	e enlightenment. The Letter to a Friend says: Without meditative concentration,	One cannot achieve wisdom awareness. On the other hand, when you have meditative	non-effect
16	acts with desire, one will be born as a hungry ghost. If	one acts with ignorance, one will be born in the animal realm. The	cause
17	as arising from the skandha - in total twenty self views. Furthermore, if	one applies the skandhas of past, present and future, there are sixty self	cause
18	skandhas of past, present and future, there are sixty self views. If	one applies the skandhas of the twenty-six directions, such as east and	cause
19	acts with hatred, one will be born in the hell realm. If	one acts with desire, one will be born as a hungry ghost. If	cause
20	of afflicting emotions, by the frequency, and by the object. First, if	one acts with hatred, one will be born in the hell realm. If	cause
21	. Relating to object, one will be born in the hell realm if	one acts nonvirtuously toward beings of higher status; if toward mediocre beings, o	cause
22	forsake all sentient beings, neither will the hawk and wolf. Therefore, if	one forsakes even one being and does not apply the antidote within a	cause
23	, taste and touch. From what causes and conditions do these arise? If	one analyses them, one ascertains that the views of worldly people, tirthikas, Vaib	cause
24	likes of Chya or Ishvara have composed various treatises explaining that if	one analyses the causes and conditions of amazing appearances – such as the brillia	cause
25	, taste and touch. From what causes and conditions do these arise? If	one analyses them, one ascertains that the views of worldly people, tirthikas, Vaib	cause
26	so on are just dream objects, only the appearance of mind. When	one analyses them through the method of the ‘one and the many’, one	cause
27	directions, such as east and so on, there are 1,560 self views. If	one applies self and other, there are 3,120 self views. Because it is the	cause
28	Son Sutra says: Anger is not the path toward enlightenment. Therefore, if	one always meditates on loving-kindness, enlightenment will be produced. II. DEFI	cause

6.3.4 The use of the generic *one* over the use of personal pronouns

The use of personal pronouns is one of the identified positive features of involved texts, as part of Biber's first dimension of the MD analysis framework (Biber, 1988; Conrad & Biber, 2001). One may argue, given that personal pronouns are indicative of involved texts, that the use of *you*, for example, in place of *one* may better fulfil the pragmatic purpose of engaging the intended audience.

Arguably, based on example 11, substituting *one* with *you,*

> "If <u>you</u> act with hatred, <u>you</u> will be reborn in the hell realm"

may resonate more strongly with an intended readership to fulfil the pragmatic purpose of religious texts to entice certain mental or physical acts than

> "If <u>one</u> acts with hatred, <u>one</u> will be reborn in the hell realm".

Yet, *one* is used in this context favour the notion of objectivity and universal applicability of the concepts communicated within the texts over the personal involvement of an intended reader, aligned with purposes of academic written texts.

> "The probable reason why generic *one* is most common in fiction and academic prose is that it is perceived as an impersonal option, lacking the personal overtones attaching to the personal pronouns when used for referring to people in general, a use more characteristic of conversation and news. […]. These are all connected with the preoccupation in academic work with making generalizations and with the wish to adopt an impersonal, objective style." (Biber et al., 1999, p. 335)

This consideration for the creation of an "impersonal, objective style" would be further upheld by the low frequency of personal pronouns, as has been seen in the wordlist chapter. In this way, the use of *one* renders the texts more objective, and as such implicitly communicates that concepts or thoughts expressed therein are "universally applicable to anyone", including the author, who, arguably could be considered to be omitted when *you* is used in the place of *one*. Similarly, when using *we*, one may consider that there is a hypothetical *they* to whom such concepts do not apply. In this way, by choosing *one* as an alternative to the personal pronouns which can imply a notion of subjectivity, and by using the more "inclusive" generic pronoun *one*, the register not only becomes more objective in its language use but at the same time removes any implicit dualisms of, for example, *we* vs. *they,* and thus the distinction between oneself (or a group I consider myself part of) and others. In this way, *one* is inclusive of all.

Thus, the way the language is used, the inclusive way of expressing cause and effect to encompass every being – including the author or speaker, rather than using an interactional “I-you” distinction that embeds the dualism of self and other - ultimate truth here conveyed: that ultimately there is no such distinction as the one between self or other. As such, even the way the language is used becomes a teaching construct to aid the comprehension of Buddhist thought.

The following section will unpick such underlying philosophical concepts that reflect the linguistic functions and uses, further.

6.4 Implications of the use of *one* on Buddhist thought

The use and function of the generic pronoun *one* that has been provided in previous sections of this chapter will be furthered in this section through the provision of an interpretation based on the philosophical thought inherent within the texts within the MDSTB corpus, and in this way illustrate that the language use is reflecting the Buddhist concepts communicated through them, namely the understanding of the nature of reality.

The use of the generic pronoun *one* in the context of Buddhism serves the purpose of providing a construct to help the reader understand the misconception of a dualism between oneself and others (Rowe, 2012, p. 21). *One* in this context is not limited to the representation of humans but instead considers *all sentient beings.* All three components of this trigram are positive keykeywords in the corpus, as has been indicated in the wordlist chapter, illustrating the centrality of this concept throughout the corpus. Such representation suggests a destabilisation of a human-centred totalising discourse by attributing equal value to animals, non-human beings and human beings. It suggests the possibility of breaking down the barrier, or dualism, between self and other, which, as has been argued, would have been implicitly expressed by the use of the personal pronouns *I, you, we* or *they* instead of *one*.

This understanding of equality between all sentient beings provides the basis for gaining an understanding of what it means to be ethical and behave responsibly according to Buddhist thought. The argument that there is no distinction between the self and other implies that one is responsible not just for oneself in our own world, but also for the wider context in which our world revolves (Garfield, 1994, p. 230).

Buddhism, as systematically expounded by Nagarjuna in the second century (Streng & Nāgārjuna, 1967, p. 32), is a logical engagement with the nature of reality. It analyses perceived phenomena ("phenomenal arising") and questions the values we ascribe to things. Its aim is to uncover the truth about reality (Streng & Nāgārjuna, 1967, p. 32), which is the literal translation of the Sanskrit word *dharma*[39], which is also the name for all teachings of the Buddha (Powers, 2010). Positive mental acts have been identified in the previous section of this chapter as the categories that lead the generic *one* to experience a positive effect, and highlight this centrality of reflection and analysis in Buddhist practice. This intention to understand the nature of reality is derived from philosophical reflection and an analysis of phenomenal arising (Samuel, 1993, p. 396).

One of the main reasons that *one* may be used so frequently here, and that there is a lack of personal pronouns in the corpus, is because Buddhists argue that the self does not exist (Bhikkhu). It is thought that we ourselves attribute meaning and value to the *I,* the self, which creates a division between self and other in that we attach to things that make us happy and repel things that cause us suffering. And it is this sense of self that prevents ethical decision-making and responsibility (Rowe, 2012, p. 115).

When making a decision, equal weight has to be given to all aspects under consideration, impartially assessing the situation to come to a just conclusion and to act accordingly. However, the self prevents this impartial view and asserts desires and aversions to create a site of privilege (Rowe, 2012).

[39] *Dharma* is a keykeyword in the MDSTB CORPUS and the most frequent loanword within the corpus. Unsurprisingly, this "logical engagement with natural phenomena", this endeavour to understand the true nature of reality, is the overarching topic of all Buddhist texts within the corpus, each text focusing on a different aspect thereof.

One has to lose one’s self-importance, let go of the self as something better and more valued than other ‘sentient beings’ (Rowe, 2012, p. 202). It is this that the corpus approach highlights. The infrequent use of *one* echoes these fundamental arguments that are at the heart of Buddhist thought and practice; the lack of a reliance on the self, and an understanding of sentient beings as having equal validity and it is a way of understanding responsibility and ethics by the way we view others.

6.5 Chapter summary

This chapter has provided an analysis of the keykeyword *one*, by drawing on word level frequency data to position as well as concordancing to examine its use and function at sentence level. Such analyses have led to the following conclusions and contributions to existing research:

This chapter has contributed to existing research by identifying the use of *one* in the context of Tibetan Buddhism as part of proper nouns (name) (e.g. *the Thus-Gone One*). Such use has been indicative of the translator’s choice to translate Sanskrit names into the English language by means of literal translation – a choice that is not persistently made by translators - thus rendering such texts incomprehensible at times to a reader who possesses only limited knowledge of the Buddhist thought. This finding furthers the characterisation of Tibetan Buddhist English as “Buddhist Hybrid English” (Griffiths, 1981), whereby the present study is based on a corpus approach rather than introspection.

The findings in this chapter have deviated from the use of *one* in academic written registers (as indicated in the LGSWE) in its use of the specific pronoun *one*. Here, *one* functions as an anaphoric reference, and in the context of Tibetan Buddhism was utilised to create cohesion at sentence level and beyond. This use of *one* is unusual for any written register as such device is most commonly associated with conversations within spoken registers. Here, the specific *one* is used to specifically create cohesion where the texts identifies a number of features of a parent category and then further elaborates on such features by means of the specific pronoun *one.* Situationally, this use has been argued to align with the Buddhist practice of memorisation[40], which utilises numbering and lists as mnemonic devices to aid memorisation of texts.

Further contributions have been made by aligning Buddhist shastras with academic written registers through the use of the generic pronoun *one.* It has been shown that *one*, in this use, most frequently functions as subject of sentence, often within the construct of a conditional subordination clause, and topically covers predominantly the subject of cause and effect. The use of *one* as the sentence subject, it has been argued, gives agency to the reader as it assumes an active function of *one* (who is associated to the cause in the “cause-effect” scenario) to achieve such effect.

Furthermore, the use of the generic *one* has been analysed in its function to express language more objectivity as compared to the use of personal pronouns that could replace *one* and as such, it has been argued, can be characterised through “universal applicability” of the content that is communicated. This depersonalisation of content, by use of the more inclusive *one*, it has been argued, aligns to Buddhist concept of the non-existence of “self” and “other”, thus the language used within the shastra becomes a vehicle for the philosophical thoughts it communicates.

[40] This concept has been dealt with only briefly in this chapter as it has been given stronger focus in chapter 7

CHAPTER 7: LOANWORDS

7.1 Introduction to the chapter

The use of loanwords in translations of shastras into Buddhist English, has been identified as one of the main reasons to render texts within this written subregister largely incomprehensible to a general audience (Griffiths, 1981, p. 20). Based on this argument, the present chapter will investigate the use of loanwords within the Mikyo Dorje Shedra of Tibetan Buddhism corpus, which is comprised of 8 shastras.

In the first part, the chapter will investigate spelling variation within the corpus, which forms one prerequisite for the corpus analysis. Such analysis will identify three categories of spelling variation, one of which relates to the transfer of loanwords and proper nouns from Sanskrit into English. Such variation is based on a non-standardised use of English within the context of Buddhism, translator's choice and sound-spelling correspondence.

The second part of the chapter will provide an analysis of loanwords based on the case study of the Sanskrit loanword *bodhicitta.* Analysis will consider

1. identification of synonyms, near synonyms and variants,

and, based on Sinclair (2004)

1. Semantic prosody
2. Collocation
3. Semantic preference

The preference of *bodhicitta* to collocate with *CULTIVATE* will be investigated in detail and compared to findings from the BNC written corpus.

Furthermore, this chapter will illustrate how concordance analysis can be used to uncover meaning the loanword *bodhicitta* to a degree, yet, arguably, Griffiths' (1981) claim of the incomprehensibility of Buddhist language, based on the use of loanwords, will be maintained.

The final section of this chapter will cast light on the Buddhist practice of textual study as part of shedras in the West, and thus indicate a mitigation in the communication of Buddhist concepts that would persist if texts were studied by a non-specialist audience, without understanding of Sanskrit or Buddhist philosophy more widely.

7.2 Spelling variation

The data within the corpus showed, upon initial wordlist generation, evidence of spelling variation within the corpus. Spelling variation has been linked to subsequent issues in the analysis of corpus such as inaccuracies in the generation of word lists, keyword lists or concordances (Baron & Rayson, 2008).

Said issues are caused by variations with words that can be classed as “orthographic words”, according to MacArthur’s taxonomy (McArthur, 1999): words with spelling variations. It was considered to approach this issue by means of addenda to the lemma list rather than to standardise the spellings within the corpus using spelling variation tools such as *Vard 2* (Baron & Rayson, 2008). The benefit of this approach is that it summarises frequency of spelling variations within the lemmatised word list whilst considering multiple spelling variations as representing one word. Spelling variation in the corpus could be categorised into three categories:

1. Variety of English (British or American spelling)
   e.g. *colorless, colourless; favourable, favourable; skilful, skilful; specialize, specialise*
2. Spelling mistakes in the publication
   In the case of spelling mistakes of loanwords were identified through spelling variation within a corpus text. Such spelling mistakes were corrected in the process of corpus compilation
3. Loanword and proper noun transfer into English
   Examples thereof can be seen in the classification of spelling variation of loanwords below.

The second category of spelling variation, based on printed spelling mistakes, were corrected within the corpus as they are not considered characteristic of the language use. The first category of British and American spelling preference is based on the place of publication of the respective text chosen within the corpus. This is a limitation within the corpus data but, given the sampling approach taken in this study, this is justified. Spelling variation based on British or American English usage were dealt with in the same way as the spelling variation within the loanword transfer, by adding such variations to the lemma list.

More interestingly than the first two categories of spelling variation is the third category, which identified orthographic variation among loanwords and proper nouns within this corpus. Such orthographical differences could be classed into further categories:

Classification of spelling variation within loanwords and proper nouns, only considering categories with a minimum frequency of 5 occurrences per category, with raw frequencies identified in “()”:

a. -amkara/-ankara
abhisamayalamkara (3), abhisamayalankara (7); sutralamkara (11), sutralankara (10); mahayanasutralamkara (3), mahayanasutralankara (2)

b. ar/ara
acharaya (4), acharya (32); arhat (23), arahats(1), arharts (1), arhats (23)

c. s/sh
ajatsatru (1), ajatashatru (5); atisa (21), atisha (6); avalokiteshvara (4), avalokitesvara (3); bhikshu (5), bhiksu (11); bhikshuni (3), bhiksuni (1); manjushri (12), manjusri (3); purusa (9), purusha (2); shastra (44), shastras (13), sastra (1); vaibhashika (5), vaibashikas (10), vaibhasika (2), vaibasikas (1)

d. v/w
bhagavan (8), bhagawan (1); ishvara (10), ishwara (2)

e. c/ch
bodhichitta (34), bodhicitta (186); kalacakra (1), kalachakra (3), vairocana (2), vairochana (1)

f. ee/i
bodhipathpradeepam (1), bodhipathpradipam (5)

g. -/d ("silent "d")
kagyu (3), *kagyud* (7)

h. ika/aka
madhyamaka (22), *madhyamika* (9), *madhyamakas* (1), *madhyamikas* (12)

i. -/h ("silent "h")
skandha (15), *skandhas* (50), *skandas* (8); *tathagata* (116), *tathaghata* (1), *tathagatas* (25); *tathagatagarbha* (48), *tatagathagarbha* (2); *vasubandhu* (7), *vasubandu* (1), *vasubhandu* (2); *vaibhashika* (5), *vaibashikas* (10), *vaibhasika* (2), *vaibasikas* (1)

Based on the overall frequency of their occurrence, spelling variations of category c (s/sh) and the example of *bodhicitta/bodhichitta* of category e will be further investigated.

-s- (53)	**-sh- (109)**	**Sanskrit root word (IAST[41])**
ajatsatru (1)	*ajatashatru* (5)	Ajātaśatru
atisa (21)	*atisha* (6)	Atiśa
avalokitesvara (3)	*avalokiteshvara* (4)	Avalokiteśvara
bhiksu (11)	*bhikshu* (5)	Bhikṣu
bhiksuni (1)	*bhikshuni* (3)	Bhikṣuṇī
manjusri (3)	*manjushri* (12)	Mañjuśrī
purusa (9)	*purusha* (2)	Puruṣa
sastra (1)	*shastra* (44), *shastras* (13)	Śāstra
vaibhasika (2), *vaibasikas* (1)	*vaibhashika* (5), *vaibashikas* (10)	Vaibhāṣika

Table 19: Spelling variation preference using s/sh within the corpus; absolute frequency counts in "()"

[41] International Alphabet of Sanskrit Transliteration (IAST)

All items listed in the table above have the shared feature that the *s* or *sh* corresponds closely to the English phoneme /ʃ/. It is perhaps for this reason that the spelling preference in English appears to be *sh*, evident in that this spelling variation was chosen over twice as frequently as the spelling variation with *s*. There are, however, exceptions, such as in the name *atisa*, which occurs almost four times as frequently as opposed to the spelling variation *atisha.* This issue may be rooted in its Sanskrit origin, where the sound of the items listed in **Error! Reference source not found.**correspond to ś or ṣ in Sanskrit (in IAST transliteration), as highlighted in the column to the right. So it may perhaps be the choice of the translator to achieve sound-spelling correspondence in English over spelling proximity to the Sanskrit terms, and thus use the *sh* spelling over the *s* spelling.

It is important, however, to consider this data in the context of the author as there appears to be a clear preference towards either *s* or *sh* spelling, as the example of *bhiks*u* and *bhiks*u* illustrates.

At first glance, there appears to be a spelling preference with *s* in reference to the Buddhist monk *bhiksu*, whereas the Buddhist nun is more frequently spelled with *sh* as in *bhikshuni* in the corpus. Further investigation into the file view, however, illustrates that the spelling choice is not dependent on the gender of the person but rather the text in which the item is used. *Bhikshu** is exclusively used in the second corpus text, the *Jewel Ornament of Liberation*, whereas *bhiksu** spelling is used solely in the third corpus text, *A Precious Garland of the Supreme Path*.

This aspect was further followed up by conducting a concordance plot search within the corpus to identify texts using the spelling variation with *s* as listed in the table above. The data revealed that 92% of the items with this spelling variation occur in corpus text 3, *A Precious Garland of the Supreme Path*. The spread of the *sh* spelling is more evenly distributed across the corpus with 4% occurring in text 1a, 30% occurring in text 2, 1% occurring in text 3, 13% occurring in text 4, 28% occurring in text 5 and 23% occurring in text 7. Each corpus text covers a distinct subtopic of Buddhism, and these spelling variations are only evident in relation to specific terminology, this accounts for the fact that there are no occurrences or only few of these spelling variations within text 1a and 1b and text 6 of the corpus.

This short analysis of the spelling variation, presented by means of the example of *s* and *sh* evidences three aspects of spelling variation of loanwords within the corpus.

1. The choice for spelling in the English translation is not standardised within the textual register.
2. Spelling choices are based on the decision-making of the translator
3. The root language, i.e. Sanskrit appears to impact this decision-making process in that a choice has to be made between achieving sound-spelling correspondence in English and closeness to the Sanskrit transliteration system on the other hand.

A lack of standardisation in the written subregister across the different texts is evident, and choices appear to be made depending on translator's or editor's preference.

The spelling of the word *bodhic*itta*[42]*,* appears to favour the spelling that corresponds to its Sanskrit spelling *bodhicitta* with its frequency almost 4.5 times as high as the spelling variant *bodhichitta.* The latter spelling bares closer proximity to English pronunciation (bodhichitta /tʃiː/) and may be the reason for this spelling choice over the Sanskrit spelling. This spelling is predominantly used in corpus text 7, with 71% of its occurrences, and only 12% of its occurrences in text 4 and 18% in text 5.

Although significantly more frequent than the *ch* spelling, the spelling *bodhicittta* is less evenly distributed across the corpus.

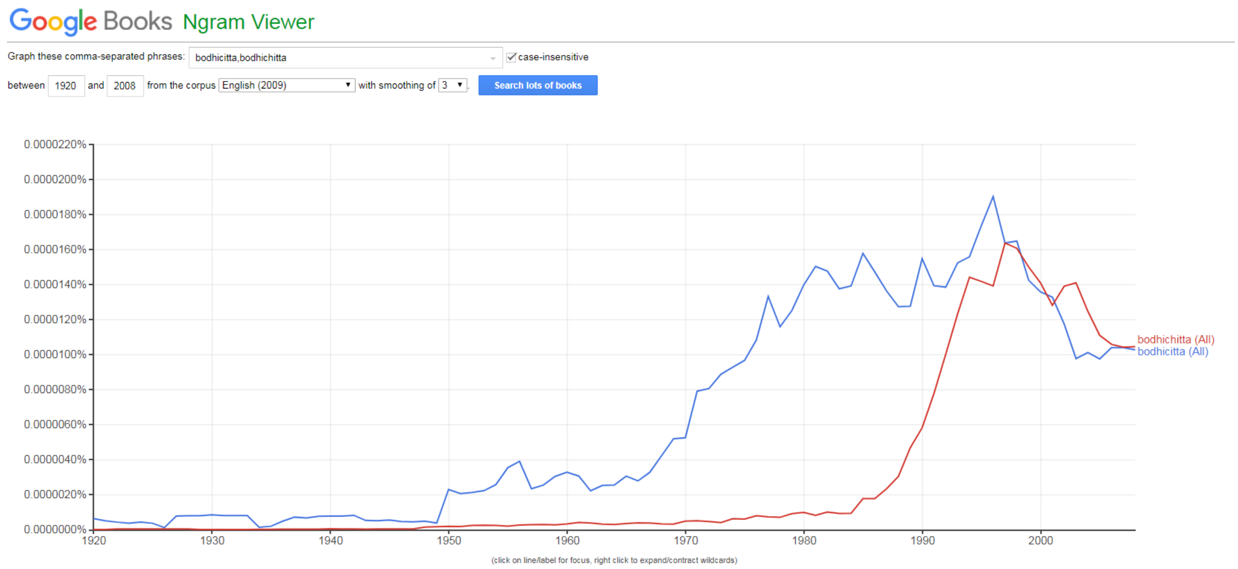
Corpus text	***bodhicitta*** % of occurrences of all spelling variations, (raw frequency)	***bodhichitta*** % of occurrences of all spelling variations, (raw frequency)
Whole corpus	84%	16%
Text 2 (US, 1998, Snow Lion)	77% (169)	
Text 3 (Taiwan, 2010)	7% (16)	

[42] *Bodhicitta* refers to the Buddhist concept of developing a mindset that strives towards enlightenment (Skt. *bodhi* means enlightened, Skt. *citta* means heart or mind)

Text 4 (UK, 1996/2015, Ganesha)		2% (4)
Text 5 (UK, 2013/2018, Ganesha)		3% (6)
Text 6 (UK, 1998, Ganesha)	0% (1)	
Text 7 (US, 2000, Snow Lion)		11% (24)

Table 20: Distribution of spelling variation of bodhicitta *and* bodhichitta *within the corpus*

The graph below illustrates that, based on the English 2009 corpus[43], historically, the spelling of *bodhicitta* is more frequent than the spelling of *bodhichitta*, which started to show an increased usage from the mid 1980s onwards, and, from the late 2000s on, both spelling variations appear to be equal in their use.



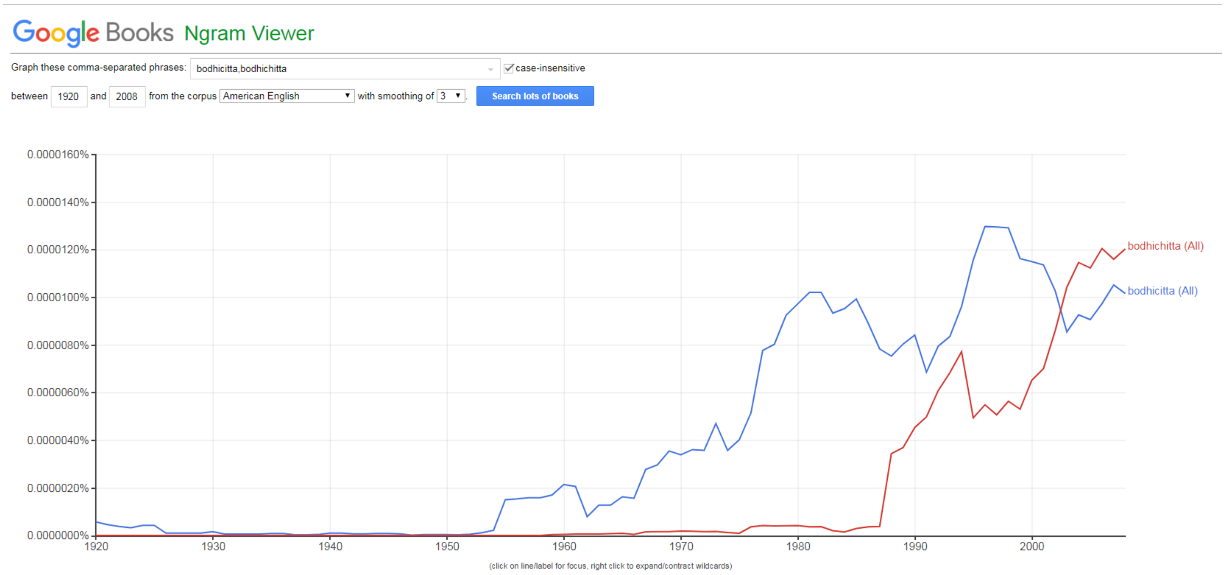
Graph these comma-separated phrases: bodhicitta,bodhichitta ☑ case-insensitive

between 1920 and 2008 from the corpus English (2009) with smoothing of 3.

Graph 4: Frequency comparision bodhicitta *and* bodhichitta *based on English 2009 Corpus. Data extracted from Google Books Ngram Viewer (Google Research, 2013a)*

Breaking down this world-wide usage of both spelling variations further illustrates a preference in the current spelling of *bochichitta* in both, the US and the UK, over a preference in the use of *bodhicitta*. The time of publication appears to be impactful in the choice of spelling with the preference towards the Sanskrit spelling up until the early 2000s in the US, and up until the late 1980s in the UK.

[43] The English 2009 Corpus is comprised of books published in the English language worldwide and digitised by Google up to the publication year 2009 (Google Research, 2013b)

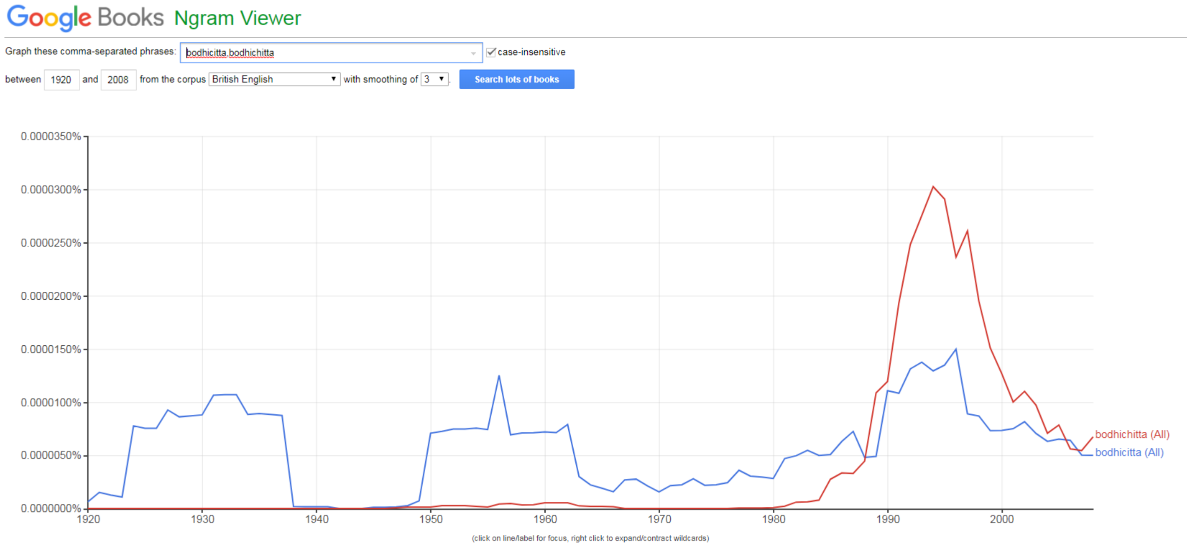


Graph these comma-separated phrases: bodhicitta,bodhichitta ☑ case-insensitive

between 1920 and 2008 from the corpus American English with smoothing of 3 Search lots of books

Graph 5: Frequency comparision bodhicitta *and* bodhichitta *based on the American English subcorpus of the English 2009 Corpus.*

Data extracted from Google Books Ngram Viewer (Google Research, 2013a)



Graph these comma-separated phrases: bodhicitta,bodhichitta ☑ case-insensitive

between 1920 and 2008 from the corpus British English with smoothing of 3 Search lots of books

Graph 6: Frequency comparision bodhicitta *and* bodhichitta *based on the British English subcorpus of the English 2009 Corpus.*

Data extracted from Google Books Ngram Viewer (Google Research, 2013a)

Conversely, this appears to have no bearing within the decision-making of the texts within the corpus, where different spelling variations are used by the same publisher, e.g. Snow Lion Publications in the US, even where the publications were produced within a 2-year period. This can similarly be observed within Ganesha Press in the

UK where different spellings are used by the same translator, though they are consistent within the individual publication.

This observation again strengthens the underlying assumption that the register indicates a lack of standardisation, here evident in the choice of spelling, even where in the wider use of the terminology beyond the corpus in published work there appear to be spelling preferences dependent both on location and date of publication.

7.3 Case study: *bodhicitta*

Bodhicitta is a frequently occurring loanword in the MDSTB corpus; it is a keyword when the corpus is compared to the BNC written. As it is key only in corpus texts 2 and 3, it is not a positive keykeyword. *Bodhicitta* has been selected for investigation due to its overall frequency within the corpus and as it is expected that the meaning of this word is largely unknown to a general audience. It is the aim of this second part of the chapter to indicate how corpus analysis can be used as a tool to uncover the meaning, use and function of *bodhicitta*, and to test if the claim of the "incomprehensibility of language" in this context still upholds (Griffiths, 1981).

Bodhicitta is a Sanskrit loanword used in the MDSTB corpus and can be literally translated into English as "awakened mind": *bodhi* (Skt.) means awakening in the sense of enlightenment, and *citta* (Skt.) means mind (Harvey, 1989).

7.3.1 (Near) synonyms of bodhicitta

In addition to the spelling variation of *bodhicitta (bodhichitta)* that has been identified in the previous section, a number of other synonyms were identified within the corpus by generating alphabetical wordlist in AntConc: *bodhicitta, bodhichitta and bodhimind.* Understanding of its literal translation (awakened mind, awakening mind, enlightened mind) informed a collocation analysis of *mind*, as well as manual analysis of concordances of such collocations, to identify any further (near-) synonyms within the corpus.

As can be seen in the table below, of *bodhicitta* and its the near-synonyms and variants, *bodhicitta* itself is the more frequent within the corpus, making up 81% of all occurrences. Within the corpus, *bodhicitta* is most frequently used in corpus text 2, accounting for 71% of all occurrences. Variants to signify the concept of *bodhicitta* are used within multiple texts within the corpus: *the developed mind* (texts 1b, 2, 3 and 7), *the cultivation of the mind* (texts 1b, 2 and 4), *enlightened mind* (texts 2, 4, 5 and 6)*, luminous (nature of) mind* (texts 3, 4, 5, 6 and 7).

The use of *bodhimind* as a near-synonym of *bodhicitta* is used only in text 1a, similarly variants, expressed by the collocations of *mind* and *awakening* (i.e. *the awakening mind)* and of *mind* and *cultivated* (i.e. *the cultivated mind* and *cultivated the mind)* are only used in text 2 to indicate *bodhicitta*.

Synonyms or near-synonymous expressions	**text 1a**	**text 1b**	**text 2**	**text 3**	**text 4**	**text 5**	**text 6**	**text 7**	**Total**
bodhicitta, bodhichitta	0	0	601	57	14	21	4	92	789
bodhimind	4	0	0	0	0	0	0	0	4
collocate of *mind*: *awakening*	0	0	14	0	0	0	0	0	14
collocate of *mind: cultivated*	0	0	25	0	0	0	0	0	25
collocate of *mind*: *cultivation*	0	4	28	0	0	0	0	4	36
collocate of *mind: developed*	0	4	4	4	0	0	0	4	14
collocate of *mind*: *enlightened*	0	0	18	0	4	4	4	0	28
collocate of *mind: luminous*	0	0	0	4	18	14	7	25	68
Total	4	7	690	64	36	39	14	124	978

Table 21: Frequency (per million tokens) of bodhicitta*, its near-synonyms and variants (based on collocates with* mind*) in the MDSTB*

Even though mere word frequency level analysis, the argument by Griffiths (1981) on the incomprehensibility of Buddhist texts, appears to uphold. Without an understanding of the literal translation of the Sanskrit term *bodhicitta*, it would be challenging to see any association between the near-synonymous or the variants listed above, all denoting the concept of *bodhicitta*. What complicates the matter further is that *bodhicitta* polysemous and represents two related, yet different concepts. Analysis of collocation patterns of *bodhicitta* will be used to identify such.

7.3.2 Collocation

Collocation has been calculated using the two-level statistical measure of mutual information (MI) and log likelihood available in AntConc (Anthony, 2019). MI identifies the collocational strength. It determines collocation by measuring “the frequency with which collocates occur together as opposed to their independent occurrence” and thus MI “will give a high collocation score to relatively low-frequency word pairs” (Baker et al., 2006, p. 38). Log likelihood (LL) provides an additional measure to test the statistical significance of the collocates by comparing the difference in observed frequencies and expected frequencies within the corpus, at a confidence level of $p<0.05$. As such log likelihood allows the researcher to rule out co-occurrence of node and collocate based on chance. The minimum frequency of a collocate with the node word (*bodhicitta*) has been set to the minimum collocate frequency of 3 with a search window span of 5L-5R in Antconc (Anthony, 2019), in order to omit any low frequency collocates from the results, and thus overcoming the limitations of the MI score. Collocates have been identified who have a minimum score of 3 (Hunston, 2002). Considering the small size of the MDSTB corpus, the two-step statistical measure of MI and LL is deemed an appropriate statistical measure to identify collocation patterns (Oakes, 1998). The collocates of the node *bodhicitta* can be seen in Appendix F: Collocates of the node *bodhicitta* (MI and LL) in MDSTB corpus.

Collocates have been filtered through manual analysis of collocates through concordances, and collocates referring to grammatical items (e.g. *the, of, in,…*) and to names of texts (e.g. *Jewel Ornament of Liberation*) have been omitted to allow a focus of the analysis on uncovering the meaning of *bodhicitta* through corpus analysis. It shall thus only be mentioned along the sidelines here that the omission included the collocate *one* which frequently collocates with *bodhicitta* in its function as a generic pronoun, and which aligns with the analysis of chapter 5, where the use of *one* in this function has been associated with the pragmatic purpose of assigning universal applicability to concepts – in this case the concept of *bodhicitta.*

7.3.3 Semantic prosody

Semantic prosody indicates, through frequent collocation, the positive or negative connotation of a node (e.g. Sinclair, 1991).

The collocational preference of the node *bodhicitta* has been investigated through identification of collocates (adjectives) from the list of collocates (see Table 22 below), and indicates a positive connotation of the loanword *bodhicitta*.

bodhicitta 222 <beneficial 12, special 5, supreme 7, perfect 3, great 4>

Such “exponents” (Stubbs, 2002p. 91), indicated in bold in the concordances below, tend to precede *bodhicitta,* as can be seen in examples 1-6 below. *Supreme* and *special* are direct modifiers of *bodhicitta*, whereas *great* and *beneficial* are part of premodified prepositional phrases: *the great power of, beneficial effects of*. The occurrence of *perfect* is, in contrast, to the right of *bodhicitta*.

1	. But one who has cultivated the **supreme**	bodhicitta enters into the Mahayana. The Bodhisattva Bhumis
2	the pleasure of peace. Cultivating the **supreme**	bodhicitta is the antidote for not understanding the
3	felt empty by the **great** power of	Bodhicitta (66) and the aspiration (67), is a quality of
4	, reveal itself!” In those beings this **special**	bodhichitta will not be born. Bodhichitta consists of
5	ment 2. **Beneficial** Effects of Cultivating Action	Bodhicitta. There are ten benefits of cultivating action
6	four causes Seeing the **beneficial** effects of	bodhicitta, Developing devotion for the Thus-gone One,
7	nament of Clear Realization says: Cultivation of	bodhicitta is the desire for **perfect**, complete enlightenment

Table 22: Concordances indicating positive discourse prosody of bodhicitta

7.3.4 Semantic preferences of *bodhicitta*[44]

This analysis has yielded a list of 64 collocates of the node *bodhicitta:*

Rank	frequency	stat	collocate
1	5	10.29951	objectives
2	3	9.34015	engendered
3	8	9.24061	lost
5	12	9.01822	cultivated
6	31	9.00894	aspiration
7	14	8.94363	cultivating
8	28	8.87804	cultivation
9	8	8.8621	developed
10	15	8.79701	cultivate
11	13	8.64827	relative
12	34	8.50026	action
13	12	8.31461	beneficial
14	7	8.28243	ceremony
15	6	8.24061	rise
16	5	8.12958	preparation
18	3	8.06004	generated
20	4	7.70456	receive
21	11	7.70004	vow
22	3	7.62394	losing
23	4	7.60834	generate
24	15	7.48075	training
26	3	7.44706	classifications
28	9	7.27303	method
31	3	7.19331	motivation
32	4	7.03898	comprise
33	3	6.93805	primary
34	6	6.93805	holding
37	5	6.73472	instructions
39	3	6.65565	obtained
40	4	6.6318	develop
41	4	6.60834	practicing
42	5	6.59907	vows
44	15	6.52106	ultimate
45	5	6.18403	special
48	6	5.78995	mahayana
50	4	5.74236	chapter
52	4	5.66772	characteristics
53	3	5.63971	realize
54	4	5.6318	given
57	7	5.42153	supreme
58	3	5.41958	times
62	3	5.31461	give
65	3	5.06004	arises
71	4	4.90433	person
72	4	4.89011	born
73	18	4.87604	has
75	8	4.83462	having
76	8	4.82107	enlightenment
78	18	4.75233	mind
80	5	4.73472	present
83	5	4.64606	bodhisattva
84	4	4.60254	bodhisattvas
88	9	4.48929	sentient
90	7	4.47205	cause
91	3	4.44014	realization
92	4	4.34337	power
93	3	4.23461	buddhas
94	4	4.21381	practice
99	3	4.06535	causes
103	24	4.01822	one
114	3	3.74236	perfect
132	6	3.3466	have
135	4	3.28409	great
136	4	3.23387	qualities
139	7	3.15073	beings

Table 23: Collocates of bodhicitta *(MI and LL) in the MDSTB corpus*

[44] To include the spelling variation *bodhichitta*, the wild card "+" was used to include both nodes in the search: *bodhic+itta (bodhicitta, bodhichitta)*

This list of collocates provides a good starting point to indicate the semantic preferences of *bodhicitta.*

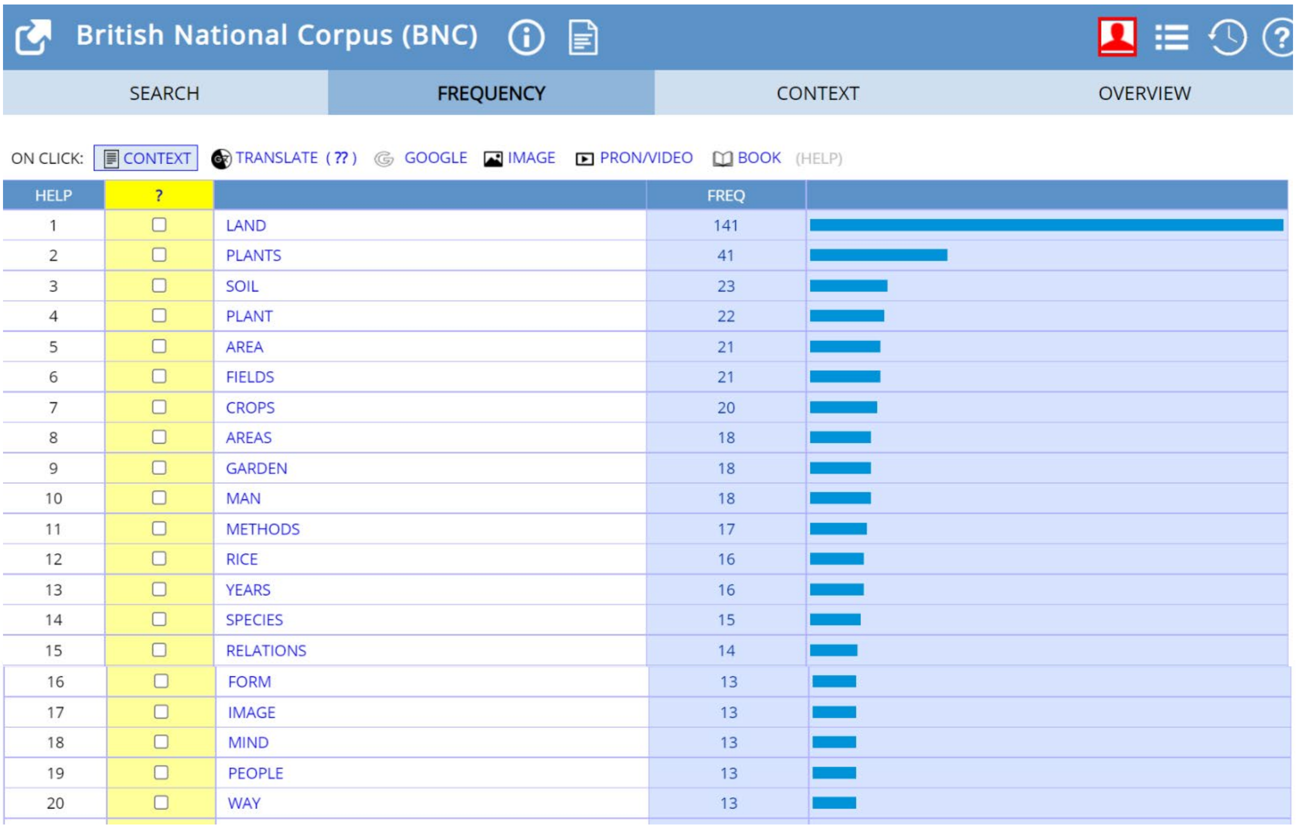
7.3.4.1 *CULTIVATE bodhicitta*

The collocates from the table above have been categorised according to their semantic proximity, all indicating how *bodhicitta* is achieved, and thus form a lexical set:

> *bodhicitta* <engendered, **cultivated**, engendered, **cultivated**, **cultivating**, **cultivation**, developed, **cultivate**, rise, generated, generate, training, obtained, develop, given, give, arises, born>

Interesting here is particularly the semantic preference of *bodhicitta* to collocate with *CULTIVATE,* as indicated in bold above. The lemma CULTIVATE collocates most frequently with *bodhicitta,* and both are collocates of one another.

In general English usage, such as indicated within the British National Corpus (BNC), CULTIVATE also shows a tendency to collocate with nouns, such as in the case of *bodhicitta.* However, its general English use indicates a strong semantic preference to collocate with the semantic field of agriculture, as can be seen in the twenty most frequent collocates of CULTIVATE in the BNC below.



HELP	?		FREQ	
1	□	LAND	141	
2	□	PLANTS	41	
3	□	SOIL	23	
4	□	PLANT	22	
5	□	AREA	21	
6	□	FIELDS	21	
7	□	CROPS	20	
8	□	AREAS	18	
9	□	GARDEN	18	
10	□	MAN	18	
11	□	METHODS	17	
12	□	RICE	16	
13	□	YEARS	16	
14	□	SPECIES	15	
15	□	RELATIONS	14	
16	□	FORM	13	
17	□	IMAGE	13	
18	□	MIND	13	
19	□	PEOPLE	13	
20	□	WAY	13	

Table 24: 20 most frequent collocates of CULTIVATE *in the BNC*

The Oxford English Dictionary (OED) defines "cultivate" (Oxford English Dictionary,) in its literal meaning as

1. "To prepare and use (land) for growing crops"
2. "To grow and improve (a plant, crop, etc.), esp. for commercial purposes."

and further in its metaphorical meaning as

3. "To refine or improve (a person, the mind, abilities, etc.) by education or training"

This use of *CULTIVATE* in the MDSTB corpus indicates a metaphorical use of *CULTIVATE,* as indicated in entry 3 of the OED.

Though infrequent, there are 13 occurrences of the collocate *MIND* with the node *CULTIVATE*, which align closely to the use in the MDSTB corpus, unsurprisingly perhaps as "mind" is the Sanskrit translation of "citta", as has been indicated above. Such instances occur exclusively in the written subcorpus: 5 occur within non-academic texts of different disciplines, 4 within religious texts, 2 within biographies and 1 in academic as well as commercial texts.

1	W_ac_humanities_arts	celebrated the mercy of a God who had granted the human **mind** sufficient illumination to **cultivate** nature and to extract those gifts necessary for subsistence. It was also a thoroughly
2	W_non_ac_humanities_arts	, are those who Leisure find, # With Care, like this, to **cultivate** their **Mind**... # ML, 2, 59. Leapor's intellectual ambitions are
3	W_non_ac_humanities_arts	ecstatic vision of God which laid aside human intelligence. It was an attempt to **cultivate** an attitude of **mind** that recognised that this intelligence -- however glorious -- had its
4	W_non_ac_medicine	can be given is,' Learn to live with it.' Try to **cultivate** an attitude of **mind** which will make it as tolerable as possible. This is
5	W_non_ac_polit_law_edu	teaching English literature came to be not the imparting of " knowledge " but the **cultivation** of the **mind**, the training of the imagination, and the quickening of the
6	W_non_ac_soc_science	there was a strong development of the sense of' culture' as the active **cultivation** of the **mind**. We can distinguish a range of meanings from (i)
7	W_religion	has been proved that creativity is a new development of the **mind** that can be **cultivated**. This faculty is widespread among the population and there has developed a system for
8	W_religion	continue to grow mentally, or he will start declining. One should aim to **cultivate** one's **mind** to its utmost potential. And this is a lifelong process;
9	W_religion	wind blowing for ever. # # -- R.S. Thomas # THE NOMADS # The **cultivation** of the **mind** entails giving it freedom to soar like a bird into the **mind**
10	W_religion	ground we are breaking. We must be our own mystics. Meditation is the **cultivation** of the **mind**; thoughts and images are the flowers; and ideas of eternal
11	W_biography	publicist of organic husbandry, Louise Matthaei Howard brought wide social sympathies, a highly **cultivated mind**, and a large measure of dry humour. She died 11 March 1969
12	W_biography	In early girlhood she vowed to overcome' the accursed thraldom of womanhood' and **cultivate** her **mind** and abilities to the utmost. She wrote this in code in her
13	W_commerce	working week too. The desire for quality is an attitude of **mind** the Profitboss **cultivates** throughout the whole organization. He'll sacrifice nothing at the expense of quality,

Table 25: Concordances of collocate MIND *with node* CULTIVATE *in the BNC*

The most frequently used pattern in the BNC *the cultivation of the mind* (concordances 5, 6, 9 and 10) resembles the two most frequent patterns of use of *bodhicitta* in the MDSTB corpus, (*the) cultivation of bodhicitta*, as can be seen in the typical phrases using the example of *CULTIVATE* in the n-grams below.

#	frequency	n-gram
1	22	cultivation of bodhicitta
2	12	the cultivation of bodhicitta
3	6	for cultivation of bodhicitta
4	4	cause for cultivation of bodhicitta
5	4	cultivate action bodhicitta
6	4	cultivate the bodhicitta
7	4	cultivating aspiration bodhicitta
8	4	to cultivate action bodhicitta
9	4	to cultivate bodhicitta
10	3	cultivating action bodhicitta
11	3	has cultivated bodhicitta
12	3	in order to cultivate action bodhicitta
13	3	of cultivating action bodhicitta
14	3	order to cultivate action bodhicitta
15	3	to cultivate the bodhicitta

Table 26: N-grams (3-6) of the collocate bodhicitta <CULTIVATE> *in the MDSTB corpus*

Examples 3 and 4 in the concordances from the BNC refer to the cultivation of *an attitude of mind,* whereby *attitude* here does not carry the common negative connotation of insolence, but rather a way of training the mind to make ones' experiences of the world more pleasant*: "Learn to live with it.' Try to <u>cultivate an attitude of mind</u> which will make it as tolerable as possible"*

This is an interesting example in that, as will be shown in the next section of this chapter, *bodhicitta* is in fact polysemous, whereby one of its frequent collocates (*aspiration)* closely relates to *attitude* in the sense that both demarcate an intentional shift in mindset.

7.3.4.2 What is *bodhicitta?*

Broadly speaking, there are two different types of bodhicitta: Relative and ultimate. Relative bodhicitta can be generated by anyone, regardless of their level or realisation, as it is the aspiration to become enlightened in order to benefit all sentient beings (rather than merely attain personal liberation). Ultimate bodhicitta correlates with a level of enlightenment that has to be achieved: the level where one realises that there is no discrimination between self and other[45]:

> "[B]odhicitta […] refers to a state of mind that corresponds to being awakened or that leads to it. It is the intention to attain perfect awakening for the sake of all beings […].
>
> According to Mahayana teachings, without this altruistic state of mind that characterizes a bodhisattva (one who has developed bodhicitta), you cannot attain the most perfect and ultimate awakening of the Buddha. […]
>
> Mahayana scriptures distinguish different aspects of bodhicitta relative to the practitioner's advancement on the path. One important distinction is that of ultimate and relative bodhicitta. […] Ultimate bodhicitta corresponds to [the] realization [that all phemomena are empty] when the mind, free from the reification of "I" and "others," is able to express true selfless compassion. Both wisdom and compassion are unified at this point.
>
> Relative bodhicitta, on the other hand, is what you can develop as an ordinary being within samsara [the cycle of rebirth and death], within a mind frame that still assumes the substantial existence of "I" and "others." This initial bodhicitta involves the wish and the commitment to attain awakening for the benefit of all beings." (Trinlay Rinpoche, 2014)

As indicated earlier, this chapter will aim to use corpus tools to aid understanding of the language used within the corpus, and in this particular example, concordances of collocates with the node *bodhicitta* have been

[45] This concept of the non-dualistic view, the non-existence of self and other has also been discussed in chapter 5 as part of the analysis of the generic pronoun *one.*

analysed (see table below) to help classify and understand the meaning of *bodhicitta*[46].

Example 1 indicates that *bodhicitta* means *Supreme Awakening*. This refers to a mental process, as indicated by example 2: the cultivation of *the highest kind of aspiration … in one's mind.*

Example 3 provides a classification of *bodhicitta* into *ultimate bodhicitta and relative bodhicitta.* Example 4 uses different terminology to denote the same concept: *not-special and special*. *Relative bodhicitta* is further subdivided into two classes in example 6: *aspiration bodhicitta* and *action bodhicitta.* Example 5 explains that *relative bodhicitta is obtained through ritual ceremony.* What such ceremony entails is further elaborated in examples 7-13. What is referred to in example 5 as *ritual ceremony* is more frequently referred to (and shows stronger collocational strength with *bodhicitta)* as *vow* (*vows).* Two different vows are the precepts for generating *bodhicitta*: 1. Taking the *bodhisattva vow* (example 7), also called *action bodhicitta vow* (example 10), which is the *vow to liberate all sentient beings from* suffering (example 8). This vow is the recitation of a *liturgy* (example 10), *three times* (example 11) A person who has taken such a vow follows the Mahayana[47] path (*enters into the Mahayana*, example 9). The second vow that is required for the generation of bodhicitta is the *pratimoksa vow*[48] (example 13), or called *pratimoksa precept* (example 12).

[46] Such classification is, of course, simplistic in nature and can by no means provide an in-depth insight into the subtle differences between the different concepts. This has, however, not been the ambition of this thesis – it ought to be considered a mere tool to aid a basic understanding through categorisation to a general audience, rather than a tool to provide a complete understanding.

[47] Mahayana and Hinayana are the two main Buddhist schools of thought. The terminology Hinayana and Mahayana is a Mahayanist way of classifying both schools. They differ in their motivation for achieving liberation (enlightenment). The Hinayanist strives for personal liberation from suffering, whereas the Mahayanist strives to reach enlightenment to liberate all being from suffering. Thus, by taking the bodhisattva vow, the vow to free all sentient beings from suffering, one becomes a Mahayana practitioner.

[48] This vow entails the adherence to a Buddhist moral conduct, which includes, for example not taking life, not lying etc.

1	ee trainings. (Trisiksa) 8. Having generated the	Bodhicitta (Supreme Awakening), perform all practice for the
2	and cultivates the highest kind of aspirations (	Bodhicitta) in one's mind and practices the
3	: There are two classes of bodhicitta: ultimate	bodhicitta and relative bodhicitta. What is ultimate bodhic
4	[the] Abhidharma: There are two types of	bodhicitta not-special and special. First, the not-
5	hrough the realization of Dharmata while relative	bodhicitta is obtained through ritual ceremony. This is
6	lassifications of relative bodhicitta: aspiration	bodhicitta and action bodhicitta. Engaging in the Conduct
7	the instructions on the development of aspiration	bodhicitta and took the bodhisattva's vow at
8	lative bodhicitta? The same sutra says: Relative	bodhicitta vows to liberate all sentient beings from
9	. But one who has cultivated the supreme	bodhicitta enters into the Mahayana. The Bodhisattva Bhumis
10	, one can receive the aspiration or action	bodhicitta vow by reciting the liturgy for either
11	space and recite the aspiration or action	bodhicitta ceremony three times and receive it that
12	pratimoksa precept in order to cultivate action	bodhicitta? It should be understood that they are
13	says that in order to cultivate action	bodhicitta, one of the pratimoksa vows is required.
14	: If one maintains the vow of action	bodhicitta And trains well in the three types
15	says: In addition, when one loses aspiration	bodhicitta, it breaks action bodhicitta The Collection of
16	is restored automatically by restoring aspiration	bodhicitta. If one broke the vow through other
17	Mahayana Sutra: At which stage does ultimate	bodhicitta arise? At the first bhumi, called Great
18	eings from suffering through compassion. Ultimate	bodhicitta is obtained through the realization of Dharmata
19	B.II.3.1.5.3. Attainment of complete perfection [	Bodhichitta] being ever-present in them the heirs
20	. Attainment of enlightenment B.II.3.1.5.2. Firm	bodhichitta B.II.3.1.5.3. Attainment of complete perfection
21	the sugatagarbha teachings. Thus the dharmakaya	bodhichitta, primordially without increase or decrease, dwell

Table 27: Concordances of collocates with the node bodhicitta

Once one has *cultivated bodhicitta,* the intention to achieve enlightenment for the benefit of all beings, the vow needs to be maintained (example 14). The loss of the vow means that one *has lost aspiration bodhicitta* and thus *broken action bodhicitta* (example 15). If the vow is maintained, one can reach *ultimate bodhicitta*. *Ultimate bodhicitta* arises *at the first bhumi*[49] (examples 16 and 17), when one has realised the *Dharmata*[50] (example 18). The following examples provide near-synonymous expressions for *ultimate bodhicitta: complete perfection* (example 19), *enlightenment* (example 20), *dharmakaya*[51] *bodhicitta* (example 21).

Based on such concordances, the following lexical set, based on the collocations of *bodhicitta* provided earlier in this chapter, is suggested to refer to the classification of *bodhicitta*:

bodhicitta <lost, cultivated, aspiration, cultivating, cultivation, developed, cultivate, relative, action, vow, losing, generate, training, classifications, motivation, comprise, vows, ultimate, special, Mahayana, characteristics, realize, supreme, arises, born, enlightenment, mind, bodhisattva, bodhisattvas, realization, practice, perfect, qualities>

Arguably, without an understanding of the target culture or Buddhist concepts, such classification would appear arbitrary and illogical. Yet, understanding the stages involved in and in relation to *bodhicitta*, will provide justification for such.

As has been shown, the use of concordances can help classify and gain a broad understanding of concepts through their representation within the corpus. The language used within this context can yet be challenging to comprehend. On the one hand, the use of synonyms, near-synonyms or variants to denote the same concept is often not obvious, and different

[49] In a very simplistic way, there are generally 10 bhumis (some classifications deviate from this number). A bhumi refers to a level of realisation (i.e. enlightenment). As such, one reaches *ultimate bodhicitta* at the first level of enlightenment.

[50] *Dharmata* (Skt.) means “the true nature of reality”.

[51] *Dharmakaya* (Skt) means “truth body” or “absolute body”

expressions to denote the same concept are not only used within the corpus but even at text level. Furthermore, the use of loanwords to explain concepts based on other loanwords, such as “Ultimate bodhicitta is obtained through the realization of Dharmata” (example 18 in the table above), clearly indicates this challenge of comprehension. Griffiths (1981) described this as the

> “tendency in contemporary Western Buddhology to retreat behind an impenetrable shield of technical vocabulary comprehensible only to co-specialists” (Griffiths, 1981, p. 20)

This finding is somewhat mitigated in the West when one considers the use of texts within Buddhist practice through shedras in the West.

7.4 Shedras in the West

A shedra,in the general sense, as has been defined in the introduction of this thesis, is a place for the study of Tibetan Buddhism, also broadly defined as “religious centres” (Phuntsho, 2000, p. 98) or “monastic universities or colleges” (Kölling, 2011, p. 18) or, more specifically, “a college of studies attached to a major monastery” (Dechen, 2008)

Shedras in the West aim at providing understanding of Buddhist philosophies to Western laypeople. As such, they may not be attached to a monastery or monastic university but to a Dharma (or Buddhist) Community. The community the shedra curriculum of the present study has been based on is the Dechen community, an “international association of Sakya and Karma Kagyu Buddhist Centres” (*Dechen.*2021). It is for this reason that the engagement with texts will be explained from the perspective of this community. Such practice, however, is aligned with the Buddhist practice of studying texts in the East.

Unlike in monastic education in the East, where the study of Buddhist texts is part of the formal education, the curriculum is spread out over a long time in the West, by providing “textual teachings”, that is lectures on Buddhist texts that form part of the shedra curriculum, twice a year[52] over about decade to complete one cycle of teachings.

Such teachings comprise the practice of what is called a *lung transmission*, whereby the text is read out aloud in its original (classical Tibetan) in full. After the lung transmission, the text is taught as part of a lecture-style delivery, whereby the lama (qualified teacher) explains first the history of the text under consideration in terms of its positioning within the wider context of literature by the author as well as its positioning within the Buddhist traditions. Following such introduction, the text is analysed sentence by sentence, and in some cases, for example for titles or headings, word for word, providing commentary and explanation of the Buddhist concepts, their etymological and socio-cultural aspects, as well as comments on (or even corrections if required) the translation. During such “lectures”, the audience (students) will be engaged by making notes throughout. After the delivery of the lecture, some time is allocated to ask any questions on the teaching (Thaye, 2013).

In the periods between the shedras, texts are further engaged with in study groups. In such small-group contexts, the students will meet, often weekly, to work through their “lecture” notes and to discuss the texts at paragraph level or at the level of one section based on the numeral structuring system (which will be analysed in chapter 7 of this thesis).

Through such study of texts much of the “fog of incomprehensibility” will be lifted, and thus safeguard the comprehension of Buddhist philosophy that may persist if texts were engaged with in isolation, without much knowledge of Sanskrit or the historical context of the texts, where, as has been shown, comprehension can easily be impeded by the use of different terminology to denote same or similar concepts (e.g. *bodhicitta, bodhimind, awakening mind, cultivation of the mind*), or by a specific loanword denoting different

[52] The shedra used to run once annually over 4 consecutive days in the winter, and since has been changed to run over two weekends twice a year (once in the summer and once in the winter)

concepts (e.g. *relative and ultimate bodhicitta*), as argued in 7.3 Case study: *bodhicitta*.

This practice of textual study as has been described in this section, arguably aligns to a large degree with the 3-step approach to the provision of translation that has been proposed by Griffiths (1981), whereby the "translator", whose role is here taken by the lama (teacher), should possess sufficient expertise to read Buddhist Sanskrit, be sufficiently well versed in the context of the text and understand the meaning of the text in order to express the author's intentions, rather than to "transfix on texts largely to transmit them by means of translation" (Griffiths, 1981, p. 20).

The Buddhist practice of studying texts, unlike Griffiths' argument, does rely on the translation of texts, whereas Griffiths, who considers the textual translations not for Buddhist practice but for academic study in the discipline of Buddhology, queries the purpose of translation of some texts altogether, as the target community of Buddhologists at large would be able to read such texts in their original, and where the original itself is considered somewhat "obscure" in its language use (Griffiths, 1981, p. 22).

In this way, following the analytical framework of Biber and Conrad (2013) provided a link between the language analysis and the situational engagement with the text by its target culture, which has been missing from previous research, where texts have been considered solely within their use as part of academic study within the disciplines of Buddhology or translation studies (Griffiths, 1981). As such, this chapter makes a valuable contribution to the methodological approaches utilised to investigate the Tibetan Buddhist written subgenre of shastra, which have been preoccupied with investigating the process of translation and ignoring the intended socio-cultural application.

Although the use of technical terms, and this is predominantly the use of Sanskrit loanwords in the written subregister of shastra, is widely documented as part of other written academic subregisters, such as in medical journal articles (e.g. Jabbour, 1997), they have not been identified as linguistic features within the MD framework. The framework uses word length as an indicator of within dimension 1 of the framework: informational production (Biber, 1988), yet, it is argued, that perhaps this aspect of the framework should be evaluated and consider specifically the use of technical terms and loanwords[53]. This study further proposes an extension of the framework to include features of Buddhist written registers that are currently not considered within such frameworks, which will further help the framework overcome its constraints in scope which is currently restricted to Western registers and ignores registers of other cultural origins.

7.5 Chapter summary

This chapter investigated the use of loanwords in the Mikyo Dorje Shedra of Tibetan Buddhism corpus. The use of Sanskrit loanwords in Buddhist English has been identified as one of the main reasons to render texts within this written subregister largely incomprehensible to a general audience (Griffiths, 1981, p. 20). This chapter indicated that, at large, this argument upholds when texts are considered in isolation, yet, within the context of the Buddhist practice of the study of such texts as part of so-called "Shedras in the West", these shortcomings are mitigated.

At its outset, the chapter investigated spelling variation within the corpus. Such analysis identified three categories of spelling variation: variation based on the variety of English used, spelling mistakes and spelling variation based on the transfer of loanwords and proper nouns (name) from Sanskrit into English. The third category indicated a lack of standardisation in the use of spelling within the context of Buddhism, translator's choice and sound-spelling correspondence from the root language Sanskrit as the three reasons for such observations.

[53] Practically, of course, this will be a challenging task as it would require the identification of such features as part of automated annotation.

The second part of the chapter provided an analysis of the Sanskrit loanword *bodhicitta.* The data indicated the use of not only the loanword *bodhicitta* to denote the Buddhist concept of the aspiration to reach enlightenment for the benefit of all beings but also the use of synonyms and near-synonyms (*bodhichitta*, *bodhimind)* and variants (e.g. *cultivation of the mind, awakening mind*).

Collocation analysis indicated a preference of *bodhicitta* to collocate with *CULTIVATE* and compared such findings to collocates of CULTIVATE within the BNC written corpus. Parallels between the uses were found between the collocation of *MIND* with CULTIVATE in the BNC written subcorpus when compared to the collocation of CULTIVATE with BODHICITTA in the MDSTB corpus, where in two instances the BNC indicated use in the context of *cultivating an attitude of mind*, a concept that very much resembles the use of *bodhicitta* in its use to demarcate an intention or aspiration.

The corpus tool concordances was used to help classify and gain a broad understanding of the concept of *bodhicitta* through its representation within the corpus. Yet, it was argued, the language use at large is challenging to comprehend, based on the use of loanwords. The main barrier to comprehension, it has been argued, is the use of different terminology to denote same or similar concepts (e.g. *bodhicitta, bodhimind, awakening mind, cultivation of the mind*), as well as specific loanwords denoting different concepts (e.g. *relative and ultimate bodhicitta*).

The final section of this chapter contextualised such findings with the Buddhist practice of textual study as part of shedras in the West, whereby texts are taught in lecture-style events, with commentary and contextual information provided, followed up by small study groups. Such practice indicated a mitigation in the communication of Buddhist concepts that would persist if texts were studied as a “stand-alone” by a non-specialist individual, without understanding of Sanskrit or Buddhist philosophy more widely.

CHAPTER 8: IMPLICATIONS

8.1 Introduction to the chapter

The present chapter illustrates the significance this thesis will make to others wishing to conduct similar or related studies. It is divided into two sections. The first section will highlight the methodological implications of this thesis by illustrating how the corpus' lexical representativeness was measured using lexical closure, and by indicating the broader implications for corpus compilation. The second part of this chapter will exemplify how corpus data can be applied pedagogically in a classroom.

8.2 Methodological implications

This section will indicate the lessons learned, and the implications of the methodology utilised in the present study, as well as make recommendations or provide templates for future research in similar contexts. The approach used within this study could also be used, for example, by educators wishing to prepare students for future study, or educators wishing to investigate the work produced by their students (e.g., assignments).

8.2.1 Corpus design

This section will illustrate the implications of this thesis with regards to corpus compilation by providing a template that may be used in future research. A detailed justification of the decision-making that underpins this template can be found in Chapter 4.4: Corpus Design.

The use of this corpus design will be relevant to those wishing to conduct corpus analysis with a pedagogic application. Those applications are likely beyond the Buddhist context that this study has investigated. Two examples illustrate where this approach may be appropriate to underpin future research: (1) educators wishing to prepare students for study in higher education, for example by compiling a corpus of core reading materials of prospective programmes of study (e.g. as found in reading lists); (2) educators wishing to compile a corpus of student writing to investigate under- or overused language in comparison to published work.

8.2.1.1 Sample

The sample[54] in the present study is based on external criteria and as such the contents were included based on their communicative function. This means that texts were not included or omitted due to internal criteria, i.e. the language contained therein. The sampling frame was based on a discourse community approach, which limited researcher bias in the selection of texts. As such, the decision-making over which texts to include in the corpus was based on the discourse community that uses such texts. In the example of the present study, it was decided to use the Buddhist genre of *shastra* as this genre functions much like the genre of compendium in the West. It draws together Buddhist teachings on a specific subtopic of Buddhist philosophy. As such it was well-placed to meet the purpose of the present study. Furthermore, the sample was based on the *shedra* curriculum, the monastic curriculum used to teach Buddhist thought to monks and nuns.

Applying this approach to other contexts, a sample might be based on all texts produced by a class, or all texts included within a reading list.

Furthermore, full texts were selected for inclusion in the corpus rather than text samples, as proposed in the wider literature (Connor & Upton, 2004; Flowerdew, 2004; Gesuato, 2011; Sinclair, 1995).

Basing the text selection on a discourse community approach will furthermore determine both the number of texts to be included in the corpus (e.g. all texts within a reading list, all essays produced by a class) and thus also the corpus size. This may impact the representativeness of the corpus. To measure the representativeness of the corpus, lexical closure was used in the present study. A further explanation and a template for this is provided

[54] An elaboration of the considerations underpinning the sampling approach has been provided in Chapter 4.4.3 Sample

in the next section: 8.2.2 Lexical closure as a measure for representativeness.

The table below provides a template based on the corpus design of the present study.

Text selection	External criteria discourse community approach
Number of texts	Determined by discourse community
Text length	Full text
Corpus size	No minimum size; determined by discourse community

Table 28: Corpus design template

Further to the corpus design considerations above, the corpus architecture framework by Bowker and Pearson (2002) was used in this study to provide further contextual information about the corpus, and has been provided in the table below, with the present study used as an example.

	Example
Corpus size	*281,290*
Number of texts	*8*
Medium	*Written publications*
Subject	*Buddhist philosophy*
Text type	*Shastra (similar to compendium)*

Language	*English, translated from classical Tibetan*
Publication date	*1997-2018*

Table 29: Corpus architecture template based on framework by Bowker & Pearson (2002)

8.2.1.2 Recording text length and metadata

In addition to making a record of the overall corpus architecture, individual text lengths that comprise the corpus were recorded. This indicated any internal imbalance within the corpus. In the context of the present study, one corpus text comprised almost 30% of the whole corpus. Particularly where the corpus tool reports absolute data, such as the tool AntConc (Anthony, 2019) that was used in the present study, further manual analysis had to be carried out to normalise some findings around the distribution of some lexical items within the corpus. For this reason, it is recommended to capture such data in a spreadsheet, where further statistical analysis can easily be carried out.

To allow the researcher to contextualise the findings from the study, and to identify any patterns in the variation of language use within the corpus, the present study recorded metadata for each text:

- Title of publication
- Author
- Translator
- First publication of the translation
- Edited publication of the translation
- Place of publication
- Publisher

The additional information a researcher chooses to record will be informed by the research that is conducted. In the context of this thesis, the researcher was interested in potential patterns based on texts that were in their second edition compared to those in their first edition (i.e. if the way language was used had changed between editions); whether there are different preferences in language use between translators, publishers or determined by the location of the publication. Other research, continuing with the example of a corpus of a reading list provided earlier, might include similar information to the information collected in the present study, but perhaps also the level of study (e.g. first year, final year). Capturing the demographic information of the author might additionally allow for an investigation of culturally inclusive reading lists.[55] In the context of a corpus of student work, one may wish to capture the first language of the student or perhaps the grade awarded to the work, just to name a few examples.

8.2.2 Lexical closure as a measure for representativeness

Lexical closure was used as a measure of lexical representativeness of a small corpus (see chapter 5). As the corpus architecture was determined by the discourse community – through inclusion of all full texts contained within a curriculum – it was deemed important to measure the lexical representativeness of the corpus to determine if the language represented within the corpus can be deemed representative of the language it aimed to represent. This was of particular concern due to the small size of the corpus and the small number of corpus texts contained therein. Lexical closure was deemed the most appropriate measure for lexical representativeness. The main barrier when applying this approach was that no definitive measure of what constitutes "closure" was provided in the literature. Terminology around "closure" uses hedging and approximations which posed the challenge to the researcher of determining a) what constitutes closure and b) observing when closure has been achieved within the corpus. The Sublanguage Corpus Analysis Toolkit – SubCAT - (Temnikova et al., 2014) explains the process of measuring lexical closure:

[55] Culturally inclusive reading lists have been comprised to represent a culturally diverse readership

> Lexical closure analysis. The lexical closure analysis algorithm detects the way that vocabulary size changes as increasingly large amounts of the corpus are observed. As tokens are observed sequentially, the number of types that those tokens represent is counted. The number of types observed is output at every 1,000 tokens (Temnikova et al., 2014, p. 4).

Yet, when it comes to defining exactly what constitutes “closure”, or to specify the algorithm to arrive at this outcome, the literature is elusive:

> General language samples will tend to show continued growth in the number of types as long as new tokens are observed – a lack of closure. Sublanguages will show a tapering off in the growth of the number of types after some number of tokens have been observed—in other words, closure (Temnikova et al., 2014, p. 5).

“Closure” here is defined as “tapering off”, in other words, a reduction in the number of new tokens added to the corpus. In this context, “closure” does not mean that no new tokens are added to the corpus as new items are still being added. Instead, it indicates that the number of new tokens added is reduced.

Teubert (1999) indicates how such closure, or “saturation” should be calculated:

> The corpus is said to be saturated at the lexical level if each addition of a new segment yields approximately the same number of new lexical items as the previous segment, i.e. when ‘the curve of lexical growth has become asymptotic’ (Teubert, 1999 cited in McEnery & Wilson, 2001).

Yet, it is not clear what “approximately” means in this context. What deviation from the number of new tokens is acceptable in a corpus that has achieved “closure” or “saturation”? Studies applying the SubCAT toolkit (Temnikova et al., 2014, p. 5) are similarly lacking a clear determination of exactly when closure is achieved:

> Overall, the curve for the BNC climbs faster and much farther and is still climbing at a fast rate after 453,377 tokens

> have been examined. In contrast, the curves for CRAFT and GENIA climb more slowly, climb much less, and by the time about 50,000 tokens have been examined the rate of increase is much smaller (Irina P Temnikova & K Bretonnel Cohen, 2013, p. 76).

Closure properties are measured visually by comparing graphs (see section 5.3.2 Lexical closure properties of the MDSTB corpus compared to other registers), and there is vagueness about when closure, or saturation, is achieved. It merely indicates where the rate of growth is smaller than the observed growth rate in other corpora. The present study utilised this visual approach by plotting the closure properties of the MDSTBC against other corpora, and it was visually easily apparent that the graph tapered off significantly earlier than those of other corpora where researchers had made claims of closure. This approach had been somewhat unsatisfactory as it was lacking precision but served the purpose for the present study of measuring lexical representativeness.

Cross-referencing the data from previously published and claims that were made about when corpora under investigation reached closure seems to correspond to the data point in the MDSTBC where for every new 1,000 tokens added to the corpus, 30 or fewer unique types are added. At this point, when new segments were added to the corpus, the number of new types added was “approximately the same” (Teubert, 1999) in that the observed difference in new types added from one added segment to the next did not exceed 14 and averaged 1.5. These observed measures may help future researchers to determine when lexical closure, or saturation, is achieved.

8.2.1.1 Proposed template for measuring lexical closure

1. Segmenting the corpus into 1,000-word segments
   Where the corpus is split into different corpus texts, these need to first be compiled into a single document, saved as a plain text (.txt) file. This document can then be split into 1,000-word segments using, for example, the freely available tool *AntFileSplitter* (Antony, 2017).

2. Recording data in Excel
   In a spreadsheet, set up columns to record

   a. Number of tokens
      These should list 0, 1,000, 2,000, 3,000 tokens, and so on;

   b. Number of types (absolute data)
      These can easily be calculated using corpus software, for example the freely available AntConc tool (Anthony, 2019). The number of types should be calculated after each new 1000-token segment has been added to the tool.

   c. Number of new types per added 1000-token segment
      This is the difference between the number of types (Column B in the table below) compared to the number of types after another 1,000-token segment has been added

COLUMN A	*COLUMN B*	*COLUMN C*
Number of tokens (1,000s)	**Number of types**	**Number of new types per added 1,000-token segment**
0	0	
1	376	376
2	588	212
...	...	...

Table 30: Example of a table to calculate lexical closure properties of a corpus (based on data from the MDSTB corpus)

3. Determining lexical closure, or saturation
   Once the number of new types per added 1,000-token segment (Column C above) from one added segment to the next is “approximately the same”, lexical closure, or saturation, has been achieved. To provide a reference point for future researchers, in the MDSTBC, this approximation was set to 14 as the maximum observed difference, with no more than 40 new types added per 1,000-token segment.

8.2.3 Data analysis

The data analysis approach used in this thesis may have implications for researchers wishing to conduct similar or related analyses. The overall data analysis was based on Biber and Conrad's analytic framework, and considered situational analysis, linguistic analysis and functional analysis (Biber & Conrad, 2013).

8.2.3.1 Template for analysis, based on Biber & Conrad (2013)

Situational analysis

Analysis of extralinguistic features, such as the context of the text(s) within the corpus. This may also include information about the purpose of a text or how it is used within the target community. Situational analysis also includes the metadata collected about a corpus as proposed in the corpus design template above.

Linguistic analysis[56]

Analysis to indicate language features of the corpus. Analysis can be conducted through different analytical functions.

- word frequency lists
- lemmatised frequency lists[57]
- keyword lists[58]
- keykeywordlists

[56] Linguistic analysis is conducted using corpus tools. Some tools are commercial and require a license. This thesis used the freely available AntcConc (Anthony, 2004; Anthony, 2019)

[57] E.g. using the freely available AntBNC Lemmalist https://www.laurenceanthony.net/resources/wordlists/antbnc_lemmas_ver_004.zip

[58] A number of reference corpora are freely available on https://www.laurenceanthony.net/software/antconc/.

The different wordlists provided quantitative data of the corpus that can then be followed up by further investigation into, for example

- dispersion (i.e. frequency distribution of a word across the different texts within a corpus). Dispersion can indicate if a word is, for example, only used within a particular text, or in a particular section within a text.
- Collocation (does a word frequently co-occur with another word)
- Concordances (to investigate the context in which a word appears; concordances were used in the present study to highlight polysemy of a word, for example, or semantic preferences)

Functional analysis

The functional analysis investigates the link between the situational and linguistic analysis. It can help answer questions such as: How does the situation in which a text occurs inform the language used therein? How does it help fulfil the communicative purpose of the text?

Studies attempting to position text types alongside other genres or communicative functions may use use the dimensions of Biber's multi-dimensional (MD) analysis (Biber, 1988) to inform the functional analysis. In the present thesis, the findings from the corpus were represented within Dimension 1: Involved vs. informational production (table 4, Features of patterns, p. 75).Whilst the use of the MD analysis framework is proposed, the researcher should remain critical of the functions associated with linguistic findings. The present study challenged some of the associated features in dimension 1, e.g. type-token ratio linked to involved texts and the proposed inclusion of technical terms in the dimension as a positive feature of informational production (section 5.4, Lexical repetition).

Furthermore, the LGSWE (Biber et al., 1999) allows for further positioning of text types alongside other registers.

This approach of combining situational, linguistic and functional analysis (Biber & Conrad, 2013) is particularly helpful when recontextualising linguistic findings with their communicative functions as well as authentic uses or practices around texts. In the present thesis, this was presented in the example of the practices of memorisation and debate within the Buddhist monastic curriculum (section 5.4., Lexical repetition), just to name one.

8.3 Pedagogic implications

The corpus design template provided in the previous section of this chapter has been created with a pedagogic application in mind. This second section of the implications chapter will provide some examples of how such corpus data may be applied in a classroom. The principles of the "Corpus Aided Platform for Language Teachers" (The Education University of Hong Kong, 2022, para. 7) and the work by Brian Tomlinson (2011) have informed the formulation of principles underpinning the example materials in this section. These principles may be utilised by others wishing to use corpus data to develop classroom materials:

1. Start with a prediction task to generate students' interest and test their prior knowledge related to the word or phrase
2. Pre-select corpus data for student exploration
3. Provide clear and simple instructions to help students achieve the tasks set.
   Where appropriate, scaffold by, for example, providing an example, signposting what needs to be done or elicit the first response in plenary before students work independently
4. Ensure tasks progress from simple to complex (in terms of cognitive engagement), and from restricted to less restricted
5. Encourage students to explore the corpus data to notice patterns and/or confirm/reject pre-existing beliefs about language use
6. To cater for students working at a different pace, include an extension activity for students who may complete tasks faster than others in class

The two sets of guided discovery classroom materials provided below focus on an investigation of Buddhist loanwords, based on the analysis of *bodhicitta* from chapter (section 7.3., Case study bodhicitta). Instructions are deliberately using simplified terminology rather than linguistic terminology (e.g. “word” rather than “lexical item” or “node”) in line with principle 3 above, “provide clear and simple instructions”.

8.3.1 Teaching types and meaning of *bodhicitta*

8.3.1.1 Prediction

(for students with prior knowledge of Buddhist philosophy)
Ask students to discuss in groups: Have you heard the term *bodhicitta* before? What do you remember about its meaning?

Or

Show students the concordance lines with the blanked out word *bodhicitta* (below). Can they guess which one word fits in all the gaps?

1	ee trainings. (Trisiksa) 8. Having generated the		(Supreme Awakening), perform all practice for the
2	and cultivates the highest kind of aspirations (		) in one’s mind and practices the
3	: There are two classes of bodhicitta: ultimate		and relative bodhicitta. What is ultimate bodhic
4	[the] Abhidharma: There are two types of		not-special and special. First, the not-
5	hrough the realization of Dharmata while relative	?	is obtained through ritual ceremony. This is
6	lassifications of relative bodhicitta: aspiration		and action bodhicitta. Engaging in the Conduct
7	the instructions on the development of aspiration		and took the bodhisattva’s vow at
8	lative bodhicitta? The same sutra says: Relative		vows to liberate all sentient beings from
9	. But one who has cultivated the supreme		enters into the Mahayana. The Bodhisattva
10	, one can receive the aspiration or action		vow by reciting the liturgy for either

Table 31: Gap fill activity: Concordances of collocates with the node *bodhicitta*

(for students with no prior knowledge of Buddhist philosophy)

Write the following statement on the board:

“*Bodhicitta* is Sanskrit for *bodhi* (awakened) + *citta* (mind)”

In groups, discuss: In the context of Buddhist thought, what do you think *bodhicitta* means? Groups should record their thoughts, for example on a whiteboard, on a flipchart or electronically on an online mind map such as Padlet (www.padlet.com).

8.3.1.2 Guided discovery: exploring types and meaning of *bodhicitta*

1	ee trainings. (Trisiksa) 8. Having generated the	Bodhicitta (Supreme Awakening), perform all practice for the
2	and cultivates the highest kind of aspirations (	Bodhicitta) in one’s mind and practices the
3	: There are two classes of bodhicitta: ultimate	bodhicitta and relative bodhicitta. What is ultimate bodhic
4	[the] Abhidharma: There are two types of	bodhicitta not-special and special. First, the not-
5	hrough the realization of Dharmata while relative	bodhicitta is obtained through ritual ceremony. This is
6	lassifications of relative bodhicitta: aspiration	bodhicitta and action bodhicitta. Engaging in the Conduct
7	the instructions on the development of aspiration	bodhicitta and took the bodhisattva’s vow at
8	lative bodhicitta? The same sutra says: Relative	bodhicitta vows to liberate all sentient beings from
9	. But one who has cultivated the supreme	bodhicitta enters into the Mahayana. The Bodhisattva Bhumis
10	, one can receive the aspiration or action	bodhicitta vow by reciting the liturgy for either
11	space and recite the aspiration or action	bodhicitta ceremony three times and receive it that
12	pratimoksa precept in order to cultivate action	bodhicitta? It should be understood that they are
13	says that in order to cultivate action	bodhicitta, one of the pratimoksa vows is required.
14	: If one maintains the vow of action	bodhicitta And trains well in the three types
15	says: In addition, when one loses aspiration	bodhicitta, it breaks action bodhicitta The Collection of
16	is restored automatically by restoring aspiration	bodhicitta. If one broke the vow through other
17	Mahayana Sutra: At which stage does ultimate	bodhicitta arise? At the first bhumi, called Great
18	eings from suffering through compassion. Ultimate	bodhicitta is obtained through the realization of Dharmata
19	B.II.3.1.5.3. Attainment of complete perfection [	Bodhichitta] being ever-present in them the heirs
20	. Attainment of enlightenment B.II.3.1.5.2. Firm	bodhichitta B.II.3.1.5.3. Attainment of complete perfection
21	the sugatagarbha teachings. Thus the dharmakaya	bodhichitta, primordially without increase or decrease, dwell

Table 32: Concordances of collocates with the node bodhicitta

Read the concordance lines provided and answer the questions below.

1. How many different types of *bodhicitta* can you identify?
2. What are they called? Highlight them in each concordance line.
3. Write a definition for each type of *bodhicitta* based on the information in the concordance lines.

8.3.1.3 Consolidation

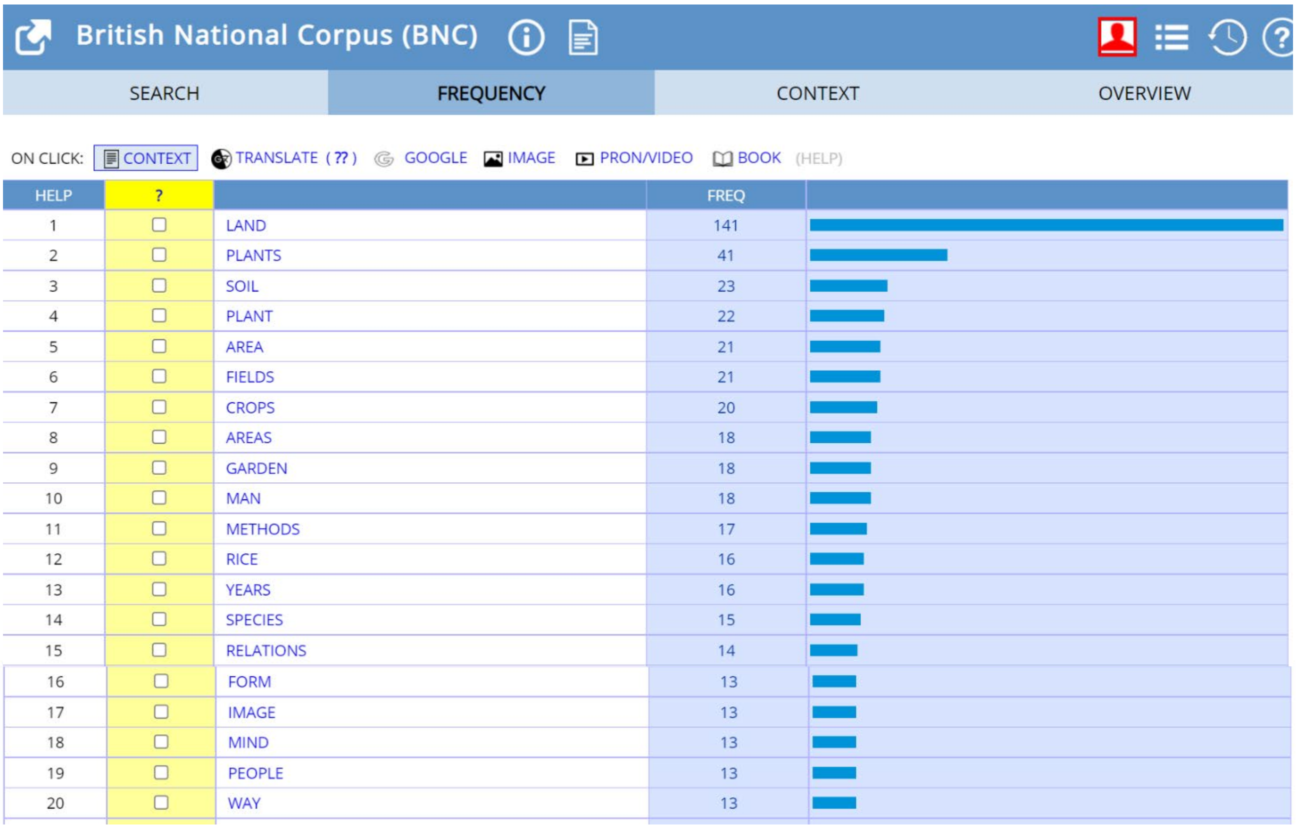
4. Compare your definition with that of another student. Agree on the best definition for each type of *bodhicitta.* You may consolidate both of your definitions
5. Discuss with a partner: What are the main differences between the different types of *bodhicitta?* What evidence can you find in the concordance lines to support this?
6. Look back at your initial predictions. What have you learned about the meaning of *bodhicitta*?

8.3.1.4 Follow-up: using bodhicitta in context

Read the concordance lines provided (table above) again and answer the questions below.

1. How do we use *bodhicitta* in a sentence? Read the three sentences below and decide, based on the concordance lines, which verbs you could use in each gap? There may be more than one possibility.
    a. If you wish to _____ action bodhicitta, you need to take the *pratimoksa vow* [the vow to reach enlightenment].
    b. You can _______ bodhicitta by taking part in a ritual ceremony.
    c. Ultimate bodhicitta is _________ at the *first bhumi* [a level of realisation].

2. In a Buddhist context, *bodhicitta* commonly collocates (=words that often co-occur) with *cultivate*.



HELP	?		FREQ	
1	□	LAND	141	
2	□	PLANTS	41	
3	□	SOIL	23	
4	□	PLANT	22	
5	□	AREA	21	
6	□	FIELDS	21	
7	□	CROPS	20	
8	□	AREAS	18	
9	□	GARDEN	18	
10	□	MAN	18	
11	□	METHODS	17	
12	□	RICE	16	
13	□	YEARS	16	
14	□	SPECIES	15	
15	□	RELATIONS	14	
16	□	FORM	13	
17	□	IMAGE	13	
18	□	MIND	13	
19	□	PEOPLE	13	
20	□	WAY	13	

Table 33: 20 most frequent collocates of CULTIVATE *in the BNC*

Look at the collocates of the word *cultivate* from the British National Corpus (see above), which shows how *cultivate* is used in a general English context

a. In the context of which topic is *cultivate* most commonly used?
b. What is the meaning of *cultivate* in this context? You can check your answers by looking up *cultivate* in a dictionary, for example the Oxford English Dictionary Online (www.oed.com).
c. Why do you think *cultivate* collocates with *bodhicitta* in a Buddhist context*?*

8.3.2 Teaching Buddhist thought through wordlists

8.3.2.1 Prediction

Think about the topic of Buddhism.

1. Which words do you think are the most frequently used ones in texts about Buddhist philosophical thought?

2. Decide on your top 5 most frequent words and put them in order of frequency.

8.3.2.2 Guided discovery

3. Compare your predictions with the list of most frequent lexical words below. How many of your words are in the list? Are they in the same order?
4. Identify one word on the list that surprises you. Investigate the word by retrieving a list of concordance lines.
    a. What can you find out about the word and how it is used in the context of Buddhism?
    b. Does the word's meaning in the context of Buddhism differ to the way you would use this word in a general English context? (teacher may wish to provide a general English corpus such as the BNC to help students)
    c. Do you think the meaning of the word in Buddhism and general English are related?

Position	Frequency per million	Word	Word forms in the corpus (and frequency)
1	4458	**buddha**	buddha 1013 buddhas 241
2	4337	**say**	said 366 say 107 saying 21 says 726
3	4238	**beings**	beings 1192
4	3861	**mind**	mind 1010 minded 3 minds 73
5	3541	**wisdom**	wisdom 936 wisdoms 60
6	3004	**other**	other 454 others 391
7	2997	**dharma**	dharma 728 dharmas 115
8	2883	**nature**	nature 805 natures 6
9	2862	**cause**	cause 477 caused 23 causes 271 causing 34
10	2862	**like**	like 801 likes 4
11	2634	**quality**	qualities 643 quality 98
12	2549	**see**	saw 13 see 251 seeing 212 seen 179 sees 62
13	2464	**great**	great 621 greater 64 greatest 8
14	2428	**path**	path 591 paths 92
15	2218	**mean**	mean 25 means 587 meant 12
16	2215	**object**	object 310 objects 313
17	2204	**way**	way 528 ways 92
18	2154	**sentient**	sentient 606
19	2144	**arise**	arise 240 arisen 36 arises 136 arising 182 arose 9
20	1962	**bodhisattava**	bodhisattava 1 bodhisattva 302 bodhisattvas 249
21	1934	**suffering**	suffering 476 sufferings 68
22	1874	**free**	free 462 freed 55 freeing 4 frees 6
23	1842	**meaning**	meaning 506 meanings 12
24	1778	**body**	bodies 64 body 436

25	1717	**practice**	practice 326 practiced 41 practices 54 practicing 62

Figure 10: Lemmatised word list of the 25 most frequent lexical items in the MDSTB corpus

8.3.2.3 Consolidation

5. Based on your research, why do you think this word is so important in Buddhist texts?

8.3.2.4 Variation: reconstructing the argument structure of a text

This task could be adapted to aid the memorisation of texts /reinforcement of arguments within texts in a Buddhist monastic context. The teacher will need to generate a word frequency list of lexical items of the text that students need to study in depth, for example in preparation for debate.

1. Present students with the title of the text the corpus is based on. Which lexical items do they expect to be highly frequent within this text?
2. Decide on your top 10 most frequent words.

Guided discovery:

3. Compare your predictions with the list of most frequent lexical words below. How many of your words are in the list? Did you add words that aren't in the wordlist?
4. Compare your own list with the corpus based wordlist and decide on a final list of the 20 most important words

Consolidation

5. Use the words in the wordlist to create a mind map of the argument within the text. You may wish to use symbols to link words, such as arrows, equal signs, etc.
6. Use your mind map to recreate the main argument of the text to a partner. Provide peer feedback: Which aspects were covered? Did you identify any gaps?
7. Work in pairs and enhance your mind maps. This will be a useful resource for your continued revision of the text.

This list of activities is indicative only. There are many other possible ways to design classroom activities based on corpus data. The main focus of the activities provided has been awareness raising. Consolidation tasks have been designed to encourage reflection on new knowledge on the one hand and to enable students to apply such learning in a meaningful way. For this reason, the consolidation have been designed are mimicking real life applications of knowledge, i.e. they are authentic. In the context of Buddhism, such authentic productive skills involve predominantly using spoken language in discussions or conversations around the topic of Buddhist philosophy. This was reflected in the tasks designed. In other contexts the design of materials to help learners develop writing skills may be more appropriate.

8.4 Chapter summary

This chapter highlighted the significance of thesis for future research. In its first part, the methodological implications of this thesis were drawn out, illustrating how the corpus' lexical representativeness was measured using lexical closure. It provided a simple template that easily allows for a replication of the approach. The first part of the chapter further indicated the broader implications for corpus compilation. The second part of this chapter highlighted the pedagogic implications and illustrated how corpus data can be applied pedagogically in a classroom.

CHAPTER 9: CONCLUSIONS

9.1 Aims and methodological approach of the study

The main aim of this study was to investigate the language used in the context of Tibetan Buddhism in order to identify the key features of the specialised domain. Such investigation was exploratory in nature, due to the lack of previous research within the field, the investigation was underpinned by the following research questions:

5. What are pervasive[59] linguistic features of the genre shastra in Tibetan Buddhist English?
6. Based on question 1, what are the characteristics of such linguistic features?
7. What is the link between such linguistic features and their situational context of Tibetan Buddhist shastras?
8. How do the linguistic features of Tibetan Buddhist shastras compare to other written registers?

In order to answer these questions, the present study applied a specialised corpus approach to the study of such language, which involved the compilation of the Mikyo Dorje Shedra of Tibetan Buddhism (MDSTB) corpus. The corpus sample was based on external criteria, namely the discourse community that uses such language and is thus comprised of full texts of the shastras that are studied as part of the Mikyo Dorje Shedra curriculum within the Karma Kagyu School of Tibetan Buddhism. As such, the 281,290 token large corpus is small in size and imbalanced in terms of text length within the corpus. Yet, such an approach allowed for the analysis of both, register and genre features. As a first data-driven linguistic analysis of the language of Tibetan Buddhism through corpus linguistics tools, the full-

[59] The term “pervasive” here is based on Biber and Conrad’s framework for register analysis, which requires “the identification of the *pervasive* linguistic features” (Biber & Conrad, 2013, p. 6)

text inclusion was deemed crucial as not to exclude pervasive features that may only occur within specific parts of the text.

Data analysis was based on Biber and Conrad’s analytic framework, and considered: situational analysis, linguistic analysis and functional analysis (Biber & Conrad, 2013). Situational analysis here denoted the analysis of extralinguistic features, such as the context of the text, and has drawn on the researcher’s own experiences or observations with the target register, on expert informants (e.g non peer reviewed material on the target culture), previous research as well as texts from the register under investigation themselves. Linguistic analysis involved the analysis of multiple texts from the same register to indicate pervasive language features of said register. The functional analysis involved the investigation of the link between the situational and linguistic analysis. The linguistic analysis identified pervasive features by means of word frequency lists, keywordlists and keykeywordlist, and further investigations considered data through frequency of distribution (dispersion), collocation and concordances, thus comparing the findings to other written (general and/or academic) registers.

Findings of the linguistic analyses were aligned with Biber’s Dimension 1, as part of his multi-dimensional (MD) analysis (Biber, 1988) or the comprehensive work of the Longman Grammar of Spoken and Written English (LGSWE) (Biber et al., 1999), which allowed the positioning of the genre of shastra alongside other written registers. Particularly the third aspect of this analytic framework, the functional analysis (i.e. the linking of situational and linguistic analysis), has been able to draw on vital contextual information about the use of texts in their Buddhist context to help understand the function of the linguistic features within the corpus. Such functional analysis has drawn on Buddhist concepts communicated through the texts as well as information about the way texts are processed within the target community, and as such go beyond.

9.2 Overview of pervasive linguistic features of Buddhist English

Data analysis within this thesis identified the following pervasive features within the MDSTB corpus. This list must, however, not be considered as exhaustive and provides an initial insight into the genre only, due to the scope of this study and the size of the corpus that such findings have been based on:

1. Vocabulary indicates tendency towards finiteness
2. High level of lexical repetition (low type-token ratio)
3. Wordlist: Most frequent tokens in the corpus reflect word frequency lists of written academic registers
    a. Low frequency of personal pronouns
    b. High frequency of prepositional postmodified noun phrases
4. High frequency of *one*
    a. Use of *one* as component of proper nouns (names)
    b. Highly frequent use of substitute pronoun *one*
    c. Highly frequent use of generic pronoun *one*
        i. Infrequently used as sentence object
        ii. Frequently used as sentence subject (most frequently in conditional subordination sentences)
5. High frequency of Sanskrit loanwords
    a. Spelling variation in the use of loanwords
    b. Use of loanword *bodhicitta*, synonym, near-synonym or variant to express Buddhist concept of
    c. Ambiguity in the use of the loanword *bodhicitta*
    d. Non-standard semantic preferences of collocates in the Buddhist context

9.3 Summary of findings and contributions made through the analysis of the Buddhist English written subregister shastra

9.3.1 Vocabulary indicates tendency towards finiteness

Lexical closure has been applied as a measure of lexical representativeness in the Mikyo Dorje Shedra of Tibetan Buddhism corpus. It was calculated by means of lexical growth analysis, and its closure properties were compared to the general written English language subcorpus of the BNC as well as the specialised written academic subregister represented in the CRAFT corpus. The MDSTB corpus indicated a tendency towards finiteness sooner than both, the general English BNC and the specialised CRAFT corpus. It has thus been illustrated that, despite the relatively small size of the specialised corpus under investigation in this study, the corpus achieved near closure and it is argued that, as such, the argument has been made that it is lexically representative of the language it aims to represents.

9.3.2 High level of lexical repetition (low type-token ratio)

Type-token ratio (TTR) has been calculated to measure the level of lexical repetition within the corpus. The MDSTB corpus indicated an unusually low TTR compared to other written registers, such as the BAWE corpus. This was an unexpected finding, particularly given the small size of the corpus. This low TTR of the corpus is caused by the high frequency of repetition of headings, subheadings and in-text table-of-contents within the corpus, indicated through analysis of full texts via fileview in Antconc. Such findings have been linked to the situational analysis of the practice of memorisation and debate in Buddhism, whereby rigorous numbering, lists and repetition of salient content through headings function as mnemonic devices in the target community.

A low type-token ratio is a linguistic feature that has previously been associated with involved production of texts as part of Dimension 1 of Biber's

MD analysis framework, and has been indicative of spoken registers (Biber, 1988; Biber, 1995; Biber et al., 1999; Reppen & Biber, 2012).

The analysis of type-token ratios in the present study, however, indicated an inconsistency to such classifications. It has been shown that the reason behind the low type-token ratio within the written subregister of shastra is not indicative of its involved production but rather is based on the frequent repetition of, for examples, headings and subheadings. Such features themselves have been found to be indicative of other academic written subregisters (Kearsey & Turner, 1999, p. 5), yet the unusually high frequency and high level of hierarchy in the use of such headings and subheadings is causing an unusually high lexical repetition and thus low type-token ratios. As such, it is suggested that the dimensions within the framework may require an extension or review based on this data.

9.3.3 Wordlist: Most frequent tokens in the corpus reflect word frequency lists of written academic registers

A word frequency data identified prevalent features of the written subregister shastra, and compared such data to the BNC written subcorpus and the BAWE corpus to position the written subregister shastra among other written registers, and to identify linguistic features within the register for further investigation, thus taking a data-driven approach.

The 20 most frequent tokens of the three corpora indicated a significant overlap between the MDSTB corpus and the written general and academic registers, whereby such similarities have been more significant between the Buddhist English register and the academic written register, only deviating in two tokens: the inclusion of *one* and *all* in the MDSTB corpus, as opposed to the inclusion of *on* and *was* in the BAWE. Such analysis indicated an alignment of the Tibetan Buddhist written subregister of shastras with academic written English registers.

9.3.3.1 High frequency of prepositional postmodified noun phrases

The most notable overlaps with general English written registers and academic written registers was the frequency of the three most frequent tokens *the, of* and *and,* indicating a frequent use of prepositional postmodified noun phrases, a feature associated with written academic language use (O'Keeffe et al., 2007). This, lead to the argument that the language use within the texts is perhaps somewhat more similar to the more formal style of writing of academic written registers than that of written English language use in general.

9.3.3.2 Low frequency of personal pronouns and high frequency of *one*

The absence of personal pronouns from the list of 20 most frequent tokens within the corpus indicates an additional linguistic features that allows the alignment of the Tibetan Buddhist written subregister of shastra alongside other written registers, in line with the identified features associated with the first dimension of Biber’s (1988) Multidimensional Analysis framework to indicate the informational (rather than involved) characteristic of the register.

9.3.4 High frequency of *one*

Three main uses of *one* have been found within the written subregister of Tibetan Buddhist shastras: (1) The use of *one* as part of a proper noun (name), (2) the use of the substitute pronoun *one*, commonly associated with spoken registers, and (3) the use of the generic pronoun *one,* frequently associated with formal written registers, yet unusually highly frequent in the present corpus as opposed to other registers.

9.3.4.1 O*ne* as acomponent of proper nouns (names)

This use of *One* as part of proper nouns (name) appears to be a feature of the specialised context of Buddhist language. *One* here is part of a compound to represent the Buddha: *the* + adjective + *One*, for example, *the Omniscient One.* Such use indicates a specific characteristic of the enlightened being. This use of proper nouns (name) indicated the challenges that the Buddhist language can pose with regards to comprehension of such texts by non-specialist audiences (Griffiths, 1981), where no explanation is provided on the meaning of such expressions in text. As such, this aspect of the study contributed to Griffiths' argument by providing empirical evidence based on data-driven research to his claims.

9.3.4.2 Highly frequent use of substitute pronoun *one*

The use of the substitute pronoun *one* was unusually high in the corpus, particularly upon comparison with the written academic subcorpus of the LGSWE corpus. Such use of the substitute *one*, commonly associated as a cohesive device in the spoken subregister of conversation due to its affordance to provide a general means of countable reference" (Biber et al., 1999, p. 334). Findings from the MDSTB corpus challenge this association which limits its association to spoken registers by highlighting its function in the Buddhist context to provide anaphoric references in texts and as contribute to text cohesion by cross-referencing to concepts that are comprised of numbered components. Such numbered lists are a common feature of the shastra, as has been indicated in the situational analysis of the Buddhist practice of memorisation where lists are one of the named mnemonic devices commonly utilised in the target culture.

9.3.4.3 Highly frequent use of generic pronoun *one*

Further contributions have been made by aligning Buddhist shastras with academic written registers through the use of the generic pronoun *one.* It has been shown that *one*, in this use, most frequently functions as subject of sentence, often within the construct of a conditional subordination clause, and topically covers predominantly the subject of cause and effect. The use of *one* as the sentence subject, it has been argued, gives agency to the reader as it assumes an active function of *one* (who is associated to the cause in the “cause-effect” scenario) to achieve a desirable effect. Such argument has also accounted for the low frequency of *one* as the subject of a sentence.

The use of the generic pronoun *one* has been analysed in its function to express language more objectivity as compared to the use of personal pronouns that could replace *one* and as such, it has been argued, can be characterised through “universal applicability” of the content that is communicated. This depersonalisation of content, by use of the more inclusive *one*, it has been argued, aligns to Buddhist concept of the non-existence of “self” and “other”, thus the language used within the shastra becomes a vehicle for the philosophical thoughts it communicates.

9.3.5 High frequency of Sanskrit loanwords

The use of Sanskrit loanwords in Buddhist English has been identified as one of the main reasons to render texts within this written subregister largely incomprehensible to a general audience (Griffiths, 1981, p. 20), due to a lack of inconsistency in their use and a lack of explanation thereof. Analysis of the MDSTB corpus identified a high frequency of loanwords within the corpus, where three of the most frequent 20 lexical items in a lemmatised wordlist referred to Sanskrit loanwords: in positions 1, 7 and 20.

9.3.5.1Spelling variation in the use of loanwords

Corpus analysis of loanwords indicated spelling variation of loanwords based on the transfer of loanwords and proper nouns (names) from Sanskrit into English. Such indicated a lack of standardisation in the use of spelling within the context of Buddhism, translator's choice and sound-spelling correspondence from the root language Sanskrit as the three reasons for such observations.

9.3.5.2 Use of loanword, synonym, near-synonym or variant to express the Buddhist concept of *bodhicitta*

Analysis of the loanword *bodhicitta,* a technical term that denotes the Buddhist concept of the aspiration to reach enlightenment for the benefit of all beings, has not been exclusively used to indicate this concept but synonyms (*bodhichitta)*, near-synonyms (*bodhimind)* and variants (e.g. *cultivation of the mind, awakening mind*) have also been used for this purpose.

This use of Sanskrit loanwords or its synonyms, near-synonyms or variants indicates the challenges that the Buddhist language poses with regards to comprehension of such texts by non-specialist audiences (Griffiths, 1981), similarly to the use of proper nouns (names) in the analysis of O*ne*. As such, this aspect of the study further supported Griffiths' claim through provision of empirical evidence.

9.3.5.3 Ambiguity in the use of the loanword *bodhicitta*

The corpus tool of concordances was used to help classify and gain a broad understanding of the concept of *bodhicitta* through its representation within the corpus. Although the use of corpus linguistics methods (concordancing) allowed the researcher to untangle such complex Buddhist concepts, and as such the methodological approach used in this study can make a valuable contribution to the study of Buddhist texts more broadly.

It was further argued, that the language use at large is challenging to comprehend for an audience not well-versed in Buddhology, based on the use of loanwords (cf. Griffiths, 1981). The main barrier to comprehension, it has been argued, is the use of different terminology to denote same or similar concepts (e.g. *bodhicitta, bodhimind, awakening mind, cultivation of the mind*), as well as specific loanwords denoting different concepts (e.g. *relative and ultimate bodhicitta*).

9.3.5.4 Mitigation of language barriers through the Buddhist practice of textual study

Provision of a situational analysis to indicate the practice of study of shastras in the West provided an insight into “Shedras in the West”, whereby texts are taught in lecture-style events, through provision of commentary and contextual information in terms of historical, socio-cultural as well as translational aspects. Such events are commonly followed up by small study groups, where the students of the Shedras engage with the text paragraph-by-paragraph, often over long periods. Such insights, it was argued, provide a mitigation in the communication of Buddhist concepts that could be misconstrued if texts were considered as a “stand-alone” and engaged with in isolation by a non-specialist individual, without understanding of Sanskrit or Buddhist philosophy more widely. Thus, applying Biber and Conrad’s (2013) analytic framework for register analyses, has provided a significant methodological contribution to previous work which failed to consider the situational contexts of such texts.

Although the use of technical terms, and this is predominantly the use of Sanskrit loanwords in the written subregister of shastra, is widely documented as part of other written academic subregisters, such as in medical journal articles (e.g. Jabbour, 1997), they have not been identified as linguistic features within the MD analysis framework. The framework uses word length as an indicator of within dimension 1 of the framework: informational production (Biber, 1988). It is for this reason, that the present study calls for an evaluation of the linguistic features contained within the Dimensions, and proposes an extension thereof to include features of Buddhist written registers that are currently not considered within such frameworks, which will further help the analysis framework overcome its constraints in scope which is currently restricted to Western registers and at large ignores registers of other cultural origins.

9.3.6 Implications of this thesis

This thesis has provided a template for other researchers wishing to conduct similar or related studies, in particular by illustrating how a corpus' lexical representativeness can be measured using lexical closure. This study has added to the existing body of research by providing a template for the calculation of lexical closure rather than by relying on visual representation. The second main contribution has been made by considering how such corpus findings can be pedagogically in a classroom. Chapter 8 was dedicated to highlighting such implications.

9.4 Limitations and further research

As the present study has been exploratory in nature, investigating a register previously untouched by corpus linguistics methods, one of the main limitations is the limitation of the linguistic features that are prevalent in English shastras of Tibetan Buddhism but that have not been considered due to the limited scope of this PhD thesis. Additionally, this thesis limited its scope to the analysis of only shastras. It would be interesting to see how the empirical insights from this study compare against the language used in other written registers within the Buddhist context, such as registers related to ritual practice (e.g. prayers) or yoga practice (e.g. *sadhanas)*, as well as spoken registers (for example based on recordings of Shedra teachings). Similarly, given that the register of shastra is still at its infancy in the English language, it would be interesting to see how the language use within the genre develops over time by means of a diachronic corpus, or to investigate translational differences of the same shastra to indicate translator's style.

Other areas for further development are mainly methodological in nature and focus on the corpus design and corpus analysis.

The Mikyo Dorje Shedra of Tibetan Buddhism corpus is small in size and, despite the argument that it has achieved lexical closure and is thus lexically representative of the language it aims to represent, a larger corpus may yield more examples of less frequent lexical items, and, more importantly, should strive for representativeness in terms of grammatical features, which has not been considered in the present study. Pervasive linguistic features within the present corpus have been shown to be over- or underrepresented within texts of the corpus, and data was further skewed by the inclusion of full texts, which created an imbalance in text size within the corpus. The use of keykeywords aimed to mitigate this shortcoming, yet the creation of a balanced corpus, based on text samples of equal length rather than full-text would help overcome such issues. In this way, a larger, more balanced corpus, would achieve a higher degree of representativeness of its language lexically and grammatically. Such a corpus could be part-of-speech tagged in order to allow for a more principled register analysis such as carrying out a multidimensional analysis (MDA) of the register, which would involve the analysis of 67 linguistic features and thus provide a more robust framework to position the written subregister of Tibetan Buddhist shastras alongside other registers, as well as follow up on the call for an evaluation of Biber's dimensions as part of the MDA made earlier in this thesis. Within the analysis of preferences of co-occurrence, a POS-tagged corpus would enable the analysis of colligation, which has been omitted from the present study. Analysis of colligation could provide interesting insights into non-standard use of words within the corpus as well as behavioural patterns of loanwords.

Arguably the most important limitation of the present study is the analysis of Buddhist texts by a non-specialist, a researcher not versed in Sanskrit or classical Tibetan, neither widely read within the body of Buddhist literature. As such, it was beyond this thesis to account for the philosophical concepts of Buddhist thought in a way that would do their nuances and complexity justice, but instead a simplistic, somewhat crude overview of the general concepts has been provided. As such, future work would benefit from a multidisciplinary approach to such investigations, drawing together expertise in empirical research and Buddhology to be able to fully use the affordances of corpus linguistics to investigate Buddhist language, particularly where the focus of such analyses is on loanwords or technical terms.

REFERENCES

Aarts, B. (2000). Corpus Linguistics, Chomsky and Fuzzy Tree Fragments. *Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20).*

Aijmer, K. (2009). *Corpora and Language Teaching*. Amsterdam: John Benjamins.

Anālayo, B. (2020). Early Buddhist Oral Transmission and the Problem of Accurate Source Monitoring. *Mindfulness, 11*(12), 2715-2724.

Anthony, L. (2004). *AntConc*. Tokyo, Japan: Waseda University.

Anthony, L. (2015). *TagAnt*. Tokyo, Japan: Waseda University.

Anthony, L. (2019). *AntConc*. Tokyo, Japan: Waseda University.

Antony, L. (2017). *File Ant Splitter (version 1.0.0)*. Tokyo, Japan: Waseda University.

Archer, D. (2009a). Does Frequency Really Matter? In D. Archer (Ed.), *What`s in a Word-list?* (pp. 1-16). London: Routledge.

Archer, D. (Ed.). (2009b). *What`s in a Word-list?* London: Routledge.

Aston, G., & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baker, P., Hardie, A., & McEnery, T. (2006). *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baron, A., & Rayson, P. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora.

Baron, A., Rayson, P., & Archer, D. (2009). Word Frequency and Key Word Statistics in Historical Corpus Linguistics. *Anglistik, 20*(1), 41.

Bhatia, V. K., Hernandez, P. S., & Perez-Paredes, P. (2011). *Researching Specialized Languages*. Amsterdam; Philadelphia: John Benjamins.

Bhatia, V., Hernández, P. S., & Pérez-Paredes, P. (2011). Specialized Languages: Corpora, meta-analyses and applications. In V. Bhatia, P. S. Hernández & P. Pérez-Paredes (Eds.), *Researching Specialized Language* (pp. 1-11). Amsterdam/London: John Benjamins.

Bhikkhu, T. Samyutta nikaya XXXV.85. *Readings in Theravada Buddhism,* Retrieved from http://www.cambodianbuddhist.org/english/website/canon/sutta/samyutta/sn35-085.html

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing, 8*(4), 243-257.

Biber, D. (1995). *Dimensions of Register Variation*. Cambridge: Cambridge University Press.

Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers*. Amsterdam, Netherlands: John Benjamins.

Biber, D., & Conrad, S. (2010). Corpus Linguistics and Grammar Teaching. Retrieved from https://www.researchgate.net/publication/265403434_Corpus_Linguistics_and_Grammar_Teaching

Biber, D., & Conrad, S. (2013). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., & Egbert, J. (2016). Register Variation on the Searchable Web. *Journal of English Linguistics, 44*(2), 95-137.

Biber, D., & Gray, B. (2011). The Historical Shift of Scientific Academic Prose in English towards less Explicit Styles of Expression. In V. Bhatia, P. S. Hernández & P. Pérez-Paredes (Eds.), *Researching Specialized Languages* (pp. 11-24). Amsterdam/Philadelphia: John Benjamins.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Fineagan, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Bingenheimer, M. (2020). *Digitization of Buddhism*. Oxford: Oxford Bibliographies.

Blumenthal, J. (2004). *The Ornament of the Middle Way: A study of the Madhyamaka thought of śāntarakita* Snow Lion Publications.

Boulton, A., Carter-Thomas, S., & Rowley-Jolivet, E. (2012). *Corpus-Informed Research and Learning in ESP : Issues and applications*. Amsterdam; Philadelphia: John Benjamins.

Bowker, L. (2000). Towards a Methodology for Exploiting Specialized Target Language Corpora as Translation Resources. *International Journal of Corpus Linguistics, 5*(1), 17-52.

Bowker, L., & Pearson, J. (2002). *Working with Specialized Language*. London: Routledge.

Breyer, Y. A. (2011). *Corpora in Language Teaching and Learning: Potential, evaluation, challenges*. Frankfurt am Main: Lang.

Campoy Cubillo, M. C., Bellés Fortuño, B., & Gea Valor, M. Luisa. (2010). *Corpus-based Approaches to English Language Teaching*. London: Continuum.

Chalmers, R. (1898). Tathāgata *The Journal of the Royal Asiatic Society, 1*, 103-115.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: De Gruyter Mouton.

Connor, U., & Upton, T. A. (Eds.). (2004). *Discourse in the Professions*. Amsterdam/Philadelphia: John Benjamins.

Connor, U., & Upton, T. A. (2004). *Applied Corpus Linguistics: A multidimensional perspective*. Amsterdam: Rodopi.

Conrad, S., & Biber, D. (Eds.). (2001). *Variation in English: Multidimensional studies*. London: Routledge.

Crystal, D. (1991). *A Dictionary of Linguistics and Phonetics* (3rd ed.). Oxford: Blackwell.

Dechen. (2008). *What is a Shedra?* Retrieved 23 February, 2014, from http://www.dechen.org/files/2012/12/Shedras.pdf

Dechen. (2021). Retrieved 12 June, 2020, from https://www.dechen.org/

Denwood, P. (1983). *Buddhist Studies*. London: Curzon Press.

Dreyfus, G. (2003). *The Sound of two Hands Clapping*. Berkeley: University of California Press.

Edgerton, F. (1953). *Buddhist Hybrid Sanskrit*. Newhaven, CN: Yale University Press.

The Eight Great Texts of the Kagyu Tradition. (n.d.). Retrieved 23 February, 2015, from http://www.nitartha.org

Fillmore, C. J. (1992). "Corpus linguistics” vs. “computer-aided armchair linguistics". *Directions in Corpus Linguistics: Proceedings from a 1991 Nobel Symposium on Corpus Linguistics* (pp. 35-66) Mouton de Gruyter.

Flowerdew, J. (2015). Corpus-based Approaches to Language Description for Specialized Academic Writing. *Language Teaching, 50*(1), 1-17.

Flowerdew, J. (2017). Corpus-based Approaches to Language Description for Specialized Academic Writing. *Language Teaching, 50*(1), 90-106.

Flowerdew, L. (2004). The Argument for Using English Specialized Corpora to Understand Academic and Professional Language. In U. Connor, & T. A. Upton (Eds.), *Discourse in the Professions: Perspectives from corpus linguistics* (pp. 11-33). Amsterdam/Philadelphia: John Benjamins.

Flowerdew, L. (2009). Applying Corpus Linguistics to Pedagogy: A critical evaluation. *International Journal of Corpus Linguistics, 14*(3), 393-417.

Flowerdew, L. (2012). *Corpora and Language Education*. New York, NY: Palgrave Macmillan.

Frankenburg-Garcia, A., Flowerdew, L., & Aston, G. (2011). *New Trends in Corpora and Language Learning*. London: Continuum.

Gaffney, S. (2000). Do the Tibetan Translations of Indian Buddhist Texts Provide Guidelines for Contemporary Translators? *SOAS Literary Review 2,* 1–15.

Garfield, J. L. (1994). Dependent Arising and the Emptiness of Emptiness: Why did Nāgārjuna start with causation? *Philosophy East & West, 44*(2), 219-250.

Garrett, E., Hill, N. W., Kilgarriff, A., Vadlapudi, R., & Zadoks, A. (2015). The Contribution of Corpus Linguistics to Lexicography and the Future of Tibetan Dictionaries. *Revue D'Etudes Tibétaines,* (32), 51-86.

Garside, R. (1987). The CLAWS Word-tagging System. In G. Leech, & G. Sampson (Eds.), *The computational analysis of English: A corpus-based approach*. London: Longman.

Gavioli, L. (2005). *Exploring Corpora for ESP learning*. Amsterdam: John Benjamins.

Gesuato, S. (2011). Structure, Content and Functions of Calls for Conference Abstracts. In V. Bhatia, P. S. Hernández & P. Pérez-Paredes (Eds.), *Discourse in the Professions: Perspectives from corpus linguistics* (pp. 87-70). Amsterdam/Philadelphia: John Benjamins.

Ghadessy, M., Henry, A., Roseberry, R. L., & Sinclair, J. (2001). *Small Corpus Studies and ELT*. Philadelphia: John Benjamins.

Google Trends. (2020). *Comparison of search terms 'gelug', 'kagyu', 'sakya' and 'nyingma' 2012-2021.* Retrieved 12 June, 2021, from https://trends.google.com/trends/explore?date=all&q=kagyu,gelug,nyingma,sakya

Griffiths, P. J. (1981). Buddhist Hybrid English: Some notes on philology and hermeneutics for Buddhologists. *The Journal of the International Association of Buddhist Studies, 4*(2), 17.

Harris, T., & Jaen, M. Moreno. (2010). *Corpus Linguistics in Language Teaching*. Oxford: Peter Lang.

Harvey, P. (1989). Consciousness Mysticism in the Discourses of the Buddha. In K. Werner (Ed.), *The Yogi and the Mystic*. London: Curzon Press.

Hoey, M. (2007). *Text, Discourse and Corpora: Theory and analysis*. London: Continuum.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2014). The contexts and Cultures of Interdisciplinary Research Discourse. *International Computer Archive of Modern and Medieval English (ICAME 35).*

Jabbour, G. (1997). *Corpus linguistics, Contextual Collocation and ESP Syllabus Creation: A text-analysis approach to the study of medical research articles.* University of Birmingham.

Jaen, M. M., Serrano, F., & Calzada Perez, M. (2010). *Exploring new Paths in Language Pedagogy: Lexis and corpus-based language teaching*. London: Equinox.

Johns, T. (2002). Data-driven Learning: The perpetual challenge. (pp. 105-117). Leiden, The Netherlands: Brill.

Kaatari, H., & Larsson, T. (2019). Using the BNC and the spoken BNC2014 to study the syntactic development of *I think* and *I'm sure*. *English Studies, 100*(6), 710-727.

Kearsey, J., & Turner, S. (1999). Evaluating Textbooks: The role of genre analysis. *Research in Science & Technological Education, 17*(1), 35-43.

Kettemann, B., & Marko, G. (2016). *Teaching and Learning by Doing Corpus Analysis: Proceedings of the fourth international conference on teaching and language corpora, Graz 19-24 July, 2000*. Leiden, The Netherlands: Brill.

Khashor, J. (2012). *Dzongsar Khyentse Chökyi Lodrö Institute: English language program curriculum.* Retrieved Jan 20, 2022, from http://jarungkhashor.blogspot.com/2011/03/dzongsar-khyentse-chokyi-lodro.html?m=0

Kölling, M. (2011). Die Renaissance des Tibetisch-Buddhistischen Klosterwesens im Kontext der Globalisierung: Ein Blick auf die soziokulturellen Wandlungsprozesse im gegenwärtigen Nepal. *Transformierte Buddhismen,* (2), 3-23.

Language Courses. (2022). Retrieved Jan 20, 2022, from https://khyentsefoundation.org/leadership-training/

Leech, G. (2011). Frequency, Corpora and Language Learning. In G. G. Meunier, J. Meunier, S. De Cock, G. Gilquin & M. Paquot (Eds.), *A taste for corpora. In honour of Sylviane Granger* (pp. 7-32). Amsterdam/Philadelphia: John Benjamins.

Lombardo, L. (2009). *Using Corpora to Learn about Language and Discourse*. Oxford: Peter Lang.

Long, L. (2013). The Translation of Sacred Texts. *The Routledge Handbook of Translation Studies* (pp. 482-492). London: Routledge.

McArthur, T. (1999). What is a Word? In T. McArthur (Ed.), *Living Words: Language, lexicography and the knowledge revolution*. Exeter: Exeter University Press.

McCarten, J., & McCarthy, M. (2010). Bridging the gap between corpus and course book: The case of conversation strategies. In A. Chambers (Ed.), *Perspectives on Language Learning Materials Development* (pp. 11-32). Oxford: Peter Lang.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction* (2nd ed.). Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

Monastery Volunteering Programme Nepal. (2022). Retrieved Jan 20, 2022, from https://www.lovevolunteers.org/destinations/volunteer-nepal/teaching-monasteries-pokhara

Nesi, H., Matheson, N., & Basturkmen, H. (2017). University Literature Essays in the UK, New Zealand and the USA: Implications for EAP. *New Zealand Studies in Applied Linguistics, 23*(2), 25-38.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Oxford English Dictionary. *"Cultivate, v.".* Retrieved Jan 20, 2018, https://www.oed.com/view/Entry/45724?redirectedFrom=cultivate&

Partington, A. (1998). *Patterns and Meanings: Using corpora for English language research and teaching*. Amsterdam: John Benjamins.

Phuntsho, K. (2000). On the two ways of learning in Bhutan. *Journal of Buthan Studies, 2*(2). 96-126.

Powers, J. (2010). *Dharma.* Retrieved 21 June, 2021, from https://www.oxfordbibliographies.com/view/document/obo-9780195393521/obo-9780195393521-0059.xml

Rayson, P. (2009). *Wmatrix: A web-based corpus processing environment*. Computing Department, Lancaster University.

Reppen, R. (2011). Using Corpora in the Language Classroom. In B. Tomlinson (Ed.), *Materials Development in Language Teaching* (pp. 35-50). Cambridge: Cambridge University Press.

Reppen, R., & Biber, D. (2012). *Corpus Linguistics*. London: SAGE.

Roiter, B. (2015). *Memorization: Beneficial Exercise for the Mind.* Retrieved Feb 18, 2019, from https://fpmt.org/mandala/online-features/memorization-beneficial-exercise-for-the-mind/

Römer, U. (2006). Pedagogical Applications of Corpora: Some reflections on the current scope and a wish list for future developments. *Zeitschrift für Anglistik und Amerikanistik, 54*(2), 121-134.

Römer, U., & Schulze, R. (2010). *Patterns, Meaningful Units and Specialized Discourses*. Amsterdam/Philadelphia, PA: John Benjamins.

Rowe, J. (2012). *Absolute Risk and the Challenge of Responsibility in 'The Dark Knight': Why the consideration of Derrida and Nagarjuna are important in a time of crisis* [PhD thesis]. Manchester Metropolitan University.

Samuel, G. (1993). *Civilized Shamans: Buddhism in Tibetan societies*. Washington DC: Smithsonian Institution Press.

Scott, M. (1997). PC Analysis of Key Words - and Key Key Words. *System, 25*(2), 233-245.

Segall, S. R. (2003). *Encountering Buddhism: Western psychology and Buddhist teachings*. Albany: State University of New York Press.

Siderits, M. (2007). *Buddhism as Philosophy*. Brookfield: Routledge.

Sinclair, J. (1991). *Corpus, Concordance, Collocation: Describing English language*. Oxford: Oxford University Press.

Sinclair, J. (1995). Corpus Typology - a framework for classification. In G. Melchers, & B. Warren (Eds.), *Studies in Anglistics* (pp. 17-33). Stockholm: Almqvist and Wiksell International.

Sinclair, J. (2004). The Search for Units of Meaning. *Trust the Text* (pp. 34-58). London: Routledge.

Sinclair, J. (2005). Corpus and Text - basic principles. In M. Wynne (Ed.), *Developing Linguistics Corpora: A guide to good practice* (pp. 1-16). Oxford: Oxbow Books.

Sinclair, J. M. (2004). *How to Use Corpora in Language Teaching*. Philadelphia: John Benjamins.

Streng, F. J., & Nāgārjuna. (1967). *Emptiness: A study in religious meaning*. Nashville, Tenn.: Abingdon Press.

Stubbs, M. (2002). *Words and Phrases* (1. publ., reprint. ed.). Oxford: Blackwell.

Sujato, B., & Brahmali, B. (2013). The Authenticity of the Early Buddhist Texts. *Journal of the Oxford Centre for Buddhist Studies, 5*.

Swales, J. (2006). *Genre Analysis: English in academic and research settings* (12. print. ed.). Cambridge: Cambridge University Press.

Teaching English to Buddhist Monks. (2022). Retrieved Jan 20, 2022, from https://www.buddhistmonasteries.org/Monastery_Projects/teaching-english-to-buddhist-monks/

Temnikova, I. P. & Cohen, K. B. (2013). Recognizing Sublanguages in Scientific Journal Articles through Closure Properties.

Temnikova, I. P., Baumgartner, J., William A, Hailu, N. D., Nikolova, I., McEnery, T., Kilgarriff, A., et al. (2014). Sublanguage corpus analysis toolkit: A tool for assessing the representativeness and sublanguage characteristics of corpora. pp. 1714-1718.

Thaye, L. J. (2001). *Way of Tibetan Buddhism*. London: Thorsons.

Thaye, L. J. (2013). Part 1 of teaching on 'one hundred pieces of advice to the people of tingri' by dampa sanjye.

The Education University of Hong Kong. (2022). *How to Create Corpus Based Materials for Classroom Use.* Retrieved Jan 20, 2022, from https://corpus.eduhk.hk/cap/how-to-create-corpus-based-materials-for-classroom-use/

Tillemans, T. J. F., & Smith, E. G. (1999). *Scripture, Logic, Language*. Somerville: Wisdom Publications.

Timmis, I. (2011). Corpora and Materials. In B. Tomlinson (Ed.), *Materials Development in Language Teaching* (pp. 461-474). London: Bloomsbury.

Timmis, I. (2013). Corpora and Materials: Towards a working relationship. In B. Tomlinson (Ed.), *Developing materials for language teaching* (pp. 461-474). London: Bloomsbury.

Timmis, I. (2015). *Corpus Linguistics for ELT*. London: Routledge.

Tomlinson Brian. (2011). *Materials Development in Language Teaching*. Cambridge: Cambridge University Press.

Tribble, C. (2000). Genres, Keywords, Teaching: Towards a pedagogic account of the language of project proposals. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 75-90). Frankfurt am Main: Peter Lang.

Tribble, C., & Jones, G. (1997). *Concordances in the classroom*. Houston: Athelstan.

Trinlay Rinpoche, K. (2014). What we've been all along: Cultivating the spirit of awakening. *Tricycle. the Buddhist Review,* Retrieved from https://tricycle.org/magazine/what-weve-been-all-along/

Upton, T. A. (2002). Understanding direct mail letters as a genre. *International Journal of Corpus Linguistics, 7*(1), 65-85.

van Auken, R. (2019, Aug 16,). Volunteer with Buddhist Monks: Teach English or work in a monastery. *Volunteer Forever.* Retrieved Jan 20, 2022, from https://www.volunteerforever.com/article_post/volunteer-with-buddhist-monks-teach-english-or-work-in-a-monastery/

Wangkhang, R. (2019, Sep 18,). Debating the Buddhist masters: A Tibetan-Canadian journalist reconnects with his culture and the dharma at a rare debate between lay practitioners and learned monastics. *Tricycle: The Buddhist Review.*

Weir, G. R. S., & Ishikawa, S. (2010). *Corpus, ICT, and Language Education*. Glasgow: University of Strathclyde Pub.

Widdowson, H. G. (1998). Context, Community and Authentic Language. *TESOL Quarterly, 32*(4), 705-716.

Williams, G. (2002). In Search of Representativity in Specialised Corpora: Categorisation through collocation. *International Journal of Corpus Linguistics, 7*(1), 43-64.

Wuthnow, R., & Cadge, W. (2004). Buddhists and Buddhism in the United States. *Journal for the Scientific Study of Religion, 43*(3), 363-380.

Xiao, R. (2010). Corpus Creation. In N. Indurkhya, & F. Damerau (Eds.), *Handbook of Natural Language Processing* (pp. 147-165). London: CRC Press.

Xiao, Z., & McEnery, A. (2005). Two Approaches to Genre Analysis. *Journal of English Linguistics, 33*(1), 62-82.

APPENDICES

Appendix A: Wordlist MDSTBC (200 most frequent items)

position	raw frequency	**frequency** (per million tokens)	item
1	25933	92193	the
2	15029	53429	of
3	9390	33382	and
4	7503	26674	is
5	5377	19116	to
6	4867	17302	in
7	3661	13015	are
8	3623	12880	a
9	3261	11593	it
10	2589	9204	that
11	2433	8649	as
12	2291	8145	by
13	2254	8013	this
14	2240	7963	one
15	2227	7917	not
16	2136	7594	all
17	2069	7355	from
18	1771	6296	be
19	1738	6179	for
20	1442	5126	with
21	1355	4817	they
22	1278	4543	so
23	1275	4533	which
24	1274	4529	or
25	1272	4522	b
26	1248	4437	there
27	1217	4326	these
28	1192	4238	beings
29	1190	4231	on
30	1106	3932	ii
31	1094	3889	will
32	1013	3601	buddha
33	1010	3591	mind
34	994	3534	three
35	968	3441	who
36	955	3395	being
37	936	3328	wisdom
38	927	3296	has
39	892	3171	have

position	raw frequency	frequency (per million tokens)	item
40	818	2908	their
41	805	2862	nature
42	801	2848	like
43	735	2613	since
44	728	2588	dharma
45	726	2581	says
46	724	2574	an
47	716	2545	when
48	701	2492	s
49	681	2421	its
50	679	2414	without
51	674	2396	i
52	670	2382	two
53	668	2375	through
54	643	2286	qualities
55	621	2208	great
56	621	2208	no
57	614	2183	those
58	606	2154	sentient
59	591	2101	he
60	591	2101	path
61	587	2087	means
62	579	2058	such
63	571	2030	his
64	567	2016	four
65	564	2005	forth
66	542	1927	if
67	528	1877	should
68	528	1877	way
69	524	1863	thus
70	514	1827	also
71	506	1799	meaning
72	477	1696	cause
73	476	1692	suffering
74	462	1642	free
75	454	1614	other
76	448	1593	explanation
77	448	1593	them
78	442	1571	primordial
79	436	1550	body
80	434	1543	any
81	432	1536	because
82	428	1522	enlightenment

position	raw frequency	frequency (per million tokens)	item
83	424	1507	having
84	422	1500	at
85	414	1472	self
86	405	1440	does
87	405	1440	what
88	396	1408	into
89	393	1397	you
90	391	1390	others
91	388	1379	mental
92	388	1379	therefore
93	386	1372	first
94	383	1362	pure
95	375	1333	time
96	373	1326	benefit
97	372	1322	five
98	366	1301	said
99	363	1290	even
100	359	1276	existence
101	354	1258	non
102	351	1248	called
103	351	1248	do
104	349	1241	sutra
105	342	1216	realms
106	339	1205	perfect
107	338	1202	can
108	326	1159	practice
109	323	1148	life
110	321	1141	but
111	319	1134	desire
112	313	1113	objects
113	310	1102	example
114	310	1102	object
115	307	1091	form
116	303	1077	dharmakaya
117	302	1074	bodhisattva
118	298	1059	power
119	284	1010	present
120	283	1006	within
121	277	985	been
122	277	985	state
123	275	978	then
124	274	974	samsara
125	272	967	six

position	raw frequency	frequency (per million tokens)	item
126	271	963	causes
127	269	956	just
128	269	956	may
129	264	939	cannot
130	264	939	compassion
131	261	928	knowledge
132	255	907	fully
133	254	903	how
134	253	899	ten
135	253	899	world
136	251	892	see
137	251	892	would
138	250	889	nirvana
139	250	889	virtue
140	249	885	bodhisattvas
141	249	885	similar
142	247	878	own
143	247	878	supreme
144	247	878	ultimate
145	246	875	explained
146	242	860	phenomena
147	241	857	buddhas
148	241	857	noble
149	240	853	arise
150	240	853	essence
151	237	843	upon
152	236	839	basis
153	234	832	teachings
154	234	832	things
155	234	832	was
156	232	825	space
157	232	825	taught
158	231	821	truth
159	230	818	we
160	229	814	consciousness
161	227	807	karma
162	225	800	types
163	222	789	tib
164	222	789	true
165	221	786	meditative
166	220	782	likewise
167	219	779	aspects
168	214	761	actions

position	raw frequency	frequency (per million tokens)	item
169	214	761	element
170	214	761	only
171	212	754	awareness
172	212	754	seeing
173	210	747	gods
174	209	743	realization
175	208	739	activity
176	208	739	result
177	208	739	understanding
178	207	736	liberation
179	204	725	born
180	203	722	emotions
181	203	722	terms
182	202	718	many
183	202	718	person
184	201	715	fruit
185	199	707	precious
186	197	700	water
187	196	697	defilements
188	193	686	completely
189	191	679	each
190	190	675	death
191	189	672	complete
192	188	668	different
193	188	668	light
194	188	668	perfection
195	187	665	meditation
196	186	661	bodhicitta
197	186	661	order
198	185	658	buddhahood
199	183	651	abandoned
200	182	647	arising

Appendix B: Lemmatised Wordlist MDSTBC (200 most frequent items)

#	raw frequency	node	lemmas
1	25471	the	the 25471
2	14645	of	of 14645
3	14480	be	am 46 are 3628 be 1753 been 274 being 931 is 7453 was 231 were 164
4	9299	and	and 9299
5	5333	to	to 5333
6	4836	in	in 4836
7	3897	it	it 3244 its 653
8	3602	a	a 3602
9	2586	they	their 798 them 445 they 1343
10	2580	that	that 2580
11	2419	as	as 2419
12	2389	one	one 2231 ones 158
13	2277	have	had 43 has 920 have 891 having 423
14	2249	by	by 2249
15	2246	this	this 2246
16	2210	not	not 2210
17	2134	all	all 2134
18	2051	from	from 2051
19	1726	for	for 1726
20	1420	with	with 1420
21	1272	or	or 1272
22	1271	so	so 1271
23	1268	which	which 1268
24	1246	buddha	buddha 1005 buddhas 241
25	1239	there	there 1239
26	1232	he	he 591 him 71 his 570
27	1220	say	said 366 say 107 saying 21 says 726
28	1206	these	these 1206
29	1190	beings	beings 1190
30	1186	on	on 1186
31	1094	will	will 1094
32	1083	mind	mind 1007 minded 3 minds 73
33	990	wisdom	wisdom 930 wisdoms 60
34	980	three	three 980
35	979	none	none 979
36	960	who	who 960
37	873	do	did 46 do 353 does 404 doing 39 done 31
38	844	other	other 454 others 390
39	836	dharma	dharma 721 dharmas 115

#	raw frequency	node	lemmas
40	812	b	b 812
41	807	i	i 622 me 53 mine 12 my 120
42	807	nature	nature 801 natures 6
43	804	like	like 800 likes 4
44	791	cause	cause 469 caused 22 causes 266 causing 34
45	735	replace	replaced 735
46	734	since	since 734
47	726	ip	ip 726
48	719	an	an 719
49	716	when	when 716
50	714	see	saw 13 see 251 seeing 209 seen 179 sees 62
51	713	quality	qualities 615 quality 98
52	694	s	s 694
53	692	great	great 620 greater 64 greatest 8
54	681	path	path 590 paths 91
55	677	without	without 677
56	657	two	two 657
57	656	through	through 656
58	653	ii	ii 653
59	622	object	object 309 objects 313
60	619	no	no 619
61	612	those	those 612
62	606	sentient	sentient 606
63	598	arise	arise 240 arisen 36 arises 136 arising 177 arose 9
64	598	mean	mean 25 means 561 meant 12
65	579	such	such 579
66	561	forth	forth 561
67	559	four	four 559
68	546	way	way 455 ways 91
69	544	bodhisattva	bodhisattava 1 bodhisattva 296 bodhisattvas 247
70	542	if	if 542
71	541	suffering	suffering 473 sufferings 68
72	533	you	you 393 your 140
73	526	should	should 526
74	523	thus	thus 523
75	520	free	free 455 freed 55 freeing 4 frees 6
76	514	also	also 514
77	483	practice	practice 326 practiced 41 practices 54 practicing 62
78	481	time	time 374 timed 1 times 106
79	478	body	bodies 64 body 414
80	471	realm	realm 129 realms 342
81	461	form	form 306 formed 7 forming 7 forms 141
82	443	benefit	benefit 368 benefiting 9 benefits 47 benefitting 19

#	raw frequency	node	lemmas
83	437	primordial	primordial 437
84	433	any	any 433
85	432	because	because 432
86	429	meaning	meaning 417 meanings 12
87	426	enlightenment	enlightenment 426
88	424	self	self 414 selves 10
89	422	at	at 422
90	413	state	state 274 stated 26 states 109 stating 4
91	410	sutra	sutra 348 sutras 62
92	398	explain	explain 32 explained 245 explaining 96 explains 25
93	394	what	what 394
94	393	into	into 393
95	390	we	our 96 us 64 we 230
96	388	therefore	therefore 388
97	387	mental	mental 387
98	387	virtue	virtue 250 virtues 137
99	386	teaching	teaching 152 teachings 234
100	385	first	first 385
101	385	give	gave 16 give 114 given 118 gives 18 giving 119
102	384	possess	possess 120 possessed 30 possesses 139 possessing 95
103	383	pure	pure 383
104	382	power	power 296 powered 1 powers 85
105	374	perfect	perfect 338 perfected 29 perfecting 6 perfects 1
106	369	existence	existence 359 existences 10
107	367	attain	attain 145 attained 137 attaining 38 attains 47
108	367	five	five 367
109	366	example	example 260 examples 106
110	364	call	call 9 called 351 calling 3 calls 1
111	364	desire	desire 319 desired 11 desires 25 desiring 9
112	364	even	even 363 evening 1
113	355	action	action 141 actions 214
114	351	non	non 351
115	348	element	element 211 elements 137
116	346	know	knew 3 know 81 knowing 94 known 121 knows 47
117	337	can	can 337
118	332	make	made 102 make 108 makes 58 making 64
119	325	result	result 208 resulting 24 results 93
120	324	present	present 282 presented 31 presenting 5 presents 6
121	323	life	life 323
122	321	but	but 321
123	319	become	became 14 become 178 becomes 91 becoming 36
124	318	take	take 122 taken 36 takes 53 taking 96 took 11
125	316	aspect	aspect 98 aspects 218

#	raw frequency	node	lemmas
126	312	truth	truth 229 truths 83
127	309	condition	condition 97 conditioned 45 conditioning 1 conditions 166
128	307	think	think 62 thinking 61 thinks 14 thought 170
129	304	appearance	appearance 172 appearances 132
130	299	dharmakaya	dharmakaya 298 dharmakayas 1
131	298	world	world 253 worlds 45
132	297	activity	activities 93 activity 204
133	294	explanation	explanation 280 explanations 14
134	288	understand	understand 141 understands 18 understood 129
135	287	thing	thing 53 things 234
136	286	consciousness	consciousness 229 consciousnesses 57
137	283	teach	taught 231 teach 28 teaches 24
138	283	within	within 283
139	281	appear	appear 149 appeared 13 appearing 38 appears 81
140	275	then	then 275
141	274	samsara	samsara 274
142	272	abandon	abandon 42 abandoned 176 abandoning 46 abandons 8
143	269	just	just 269
144	269	may	may 269
145	269	six	six 269
146	266	knowledge	knowledge 259 knowledges 7
147	264	cannot	cannot 264
148	263	god	god 53 gods 210
149	263	purify	purified 169 purifies 15 purify 43 purifying 36
150	262	compassion	compassion 262
151	261	realize	realize 88 realized 81 realizes 45 realizing 47
152	259	reason	reason 162 reasoned 1 reasons 96
153	258	bhumi	bhumi 145 bhumis 113
154	256	view	view 140 viewed 18 viewing 16 views 82
155	255	fully	fully 255
156	254	own	own 247 owned 7
157	252	type	type 29 types 222 typing 1
158	251	how	how 251
159	251	phenomenon	phenomena 241 phenomenon 10
160	251	would	would 251
161	250	ten	ten 250
162	249	exist	exist 154 existed 13 existing 23 exists 59
163	249	similar	similar 249
164	248	nirvana	nirvana 248
165	246	show	show 43 showed 2 showing 99 shown 64 shows 38
166	246	ultimate	ultimate 246
167	243	supreme	supreme 243
168	242	kaya	kaya 86 kayas 156

#	raw frequency	node	lemmas
169	242	person	person 202 persons 40
170	241	noble	noble 241
171	239	defilement	defilement 53 defilements 186
172	239	karma	karma 225 karmas 14
173	236	follow	follow 55 followed 13 following 109 follows 59
174	236	upon	upon 236
175	235	basis	basis 235
176	234	faculty	faculties 130 faculty 104
177	234	space	space 232 spaces 2
178	232	term	term 25 termed 16 terming 1 terms 190
179	231	complete	complete 185 completed 28 completes 9 completing 9
180	230	perfection	perfection 187 perfections 43
181	224	bear	bear 9 bearing 5 bears 5 born 204 borne 1
182	224	part	part 175 parting 2 parts 47
183	221	jewel	jewel 143 jewels 78
184	221	TRUE	true 221
185	220	achieve	achieve 131 achieved 56 achieves 22 achieving 11
186	220	awareness	awareness 212 awarenesses 8
187	220	bodhichitta	bodhichitta 34 bodhicitta 186
188	220	likewise	likewise 220
189	220	water	water 197 waters 23
190	218	meditative	meditative 218
191	218	wish	wish 131 wished 3 wishes 61 wishing 23
192	217	abide	abide 53 abided 1 abides 67 abiding 96
193	217	emotion	emotion 14 emotions 203
194	215	essence	essence 214 essences 1
195	214	only	only 214
196	212	experience	experience 139 experienced 26 experiences 33 experiencing 14
197	211	understanding	understanding 208 understandings 3
198	207	realization	realization 203 realizations 4
199	204	liberation	liberation 202 liberations 2
200	203	kind	kind 62 kinds 141

Appendix C: Concordances of substitute pronoun *one* in the MDSTBC

[*the* + ordinal number + *one*]

#	concordance	
1	. Of these three types of compassion, we will meditate on the first	one. B. Object. All sentient beings are its object. C. Identifying Characteristi
2	ssess four qualities, and those who possess two qualities. Concerning the first	one, Bodhisattva Bhumis says: One should understand that a bodhisattva who has eig
3	cause of losing aspiration and the cause of losing action. The first	one consists of forsaking sentient beings, adopting the four unwholesome deeds, *17
4	neself]. a) Investigating Impermanence within Oneself. Meditate on the first	one in these ways: meditate on death, meditate on the characteristics of death,
5	ence with cause and (2) interdependence supported by conditions. (1) The first	one, interior interdependence with cause. As is said: Monks, because of this, that
6	are three types of harsh words: direct, circuitous, and indirect. The first	one is forceful, directly digging at someone's various faults. The second one
7	two types of impermanence associated with the inner sentient beings, the first	one is impermanence of others. All the sentient beings in the three worlds
8	ing suffering, and patience in understanding the nature of Dharma. The first	one is practicing patience by investigating the nature of the one who creates
9	been fully actualized, they are not brought into the mind. The last	one is so named because there is very little perception; although there is
10	maturation, and the cultivation of bodhicitta with removed veils. *3 The first	one is the level of interested behavior. The second one extends from the
11	virtues— These five comprise the training in aspiration bodhicitta. The first	one is the method for not losing bodhicitta. The second one is the
12	ties, and C. meditative concentration of benefitting sentient beings. The first	one is the method to make a proper vessel of one's own
13	investigating the need to be free from it quickly. (a) The First	One, Meaninglessness. Sometimes I have committed evil deeds to subjugate enemies, s
14	own, with regard to others', and with regard to neither. *7 The first	one means being attached to one's own race, clan, body, qualities, and
15	and result, of the truth, and of the Three Jewels. The first	one means not believing that suffering and happiness are caused by nonvirtue and
16	are three types of idle talk: false, worldly, and true. The first	one means reciting the mantras and reading the texts of heretics and so
17	family, protected by the owner, and protected by the Dharma. The first	one means sexual misconduct with one's mother, sister, and so forth. *4 The
18	thought that comes from hatred, from jealousy, and from resentment. The first	one means the desire to kill others with hatred, like in a battle.
19	of hatred, and taking life through the door of ignorance. The first	one means to take life for meat, pelts and so forth, for sport,
20	by force, taking things secretly, and taking things through deceit. The first	one means to rob by force without any reason. The second one means

21	types of lying: spiritual lies, big lies, and small lies. The first	one means to lie about having a supreme Dharma quality. *6 The second one
22	ith phenomena as its object, and nonobjectified compassion. Of these, the first	one means to develop compassion by seeing the suffering of sentient beings in
23	afflicting emotions, and c) meditative concentration will arise. a) The First	One. Taking seven steps toward a monastery with the motivation to stay there
24	the completion of the act of merely snapping the fingers. The longest	one, the time from when the Buddha first made the aspiration until he
25	to make a proper vessel of one’s own mind. The second	one is establishing all of the Buddha’s qualities on the basis of
26	C. insatiable perseverance. The first is the excellent motivation, the second	one is excellent applied effort, and the third one is the perfection of
27	by investigating the nature of the one who creates harm. The second	one is practicing patience by investigating the nature of suffering. The third one
28	. The first one is the method for not losing bodhicitta. The second	one is the method by which bodhicitta does not weaken. The third one
29	one is forceful, directly digging at someone’s various faults. The second	one is to use sarcasm or some funny words just to hurt. The
30	mantras and reading the texts of heretics and so forth. The second	one is useless chatter. The third one is giving Dharma teachings to those
31	such intolerable cold that blisters cover their entire bodies. In the second	one, it is so cold the blisters burst. These names refer to the
32	, and wealth, and thinking, “There is no one like me.” The second	one means being attached to others’ prosperity and thinking, “I wish I owned
33	that suffering and happiness are caused by nonvirtue and virtue. The second	one means not believing that one attains the Truth of Cessation even if
34	sexual misconduct with one’s mother, sister, and so forth. *4 The second	one means sexual misconduct with someone owned by a husband or king, and
35	desire to kill others with hatred, like in a battle. The second	one means the desire to kill, and so forth, a competitor out of
36	’s own wealth, and to maintain oneself and loved ones. The second	one means to take life through the arising of hatred, out of resentment,
37	first one means to rob by force without any reason. The second	one means to steal things by breaking into a house without others noticing
38	one means to lie about having a supreme Dharma quality. *6 The second	one means to tell a lie that makes a difference between harm and
39	. The first means to divide friends in front of them. The second	one means to divide two friends with indirect language. The third one refers
40	first means to restrain your mind in a proper place; the second	one means to mature the Dharma qualities of your mind; and the third
41	outer fixation been shown not to have existence. Explanation of the second	one, that the mind of inner grasping is nonexistent. Some (Solitary Realisers and
42	of sentient beings in the lower realms and so forth The second	one when one is well trained in the practice of the Four Noble
43	someone owned by a husband or king, and so forth. The third	one has five subcategories: even with one’s own wife, sexual misconduct refers
44	to others’ prosperity and thinking, “I wish I owned this.” The third	one is being attached to an underground mine, or the like, which belongs
45	Buddha’s qualities on the basis of the proper vessel. The third	one is benefitting sentient beings. IV. CHARACTERISTICS OF EACH CLASSIFICATION A. A

46	heretics and so forth. The second one is useless chatter. The third	one is giving Dharma teachings to those who have no respect and who
47	, the first two are practiced in the conventional state, and the third	one is practiced according to the ultimate state. IV. CHARACTERISTICS OF EACH CLA
48	one is practicing patience by investigating the nature of suffering. The third	one is practicing patience by investigating the unmistakable nature of all phenomen
49	one is the method by which bodhicitta does not weaken. The third	one is the method for increasing the strength of bodhicitta. The fourth one
50	excellent motivation, the second one is excellent applied effort, and the third	one is the perfection of these two. IV. CHARACTERISTICS OF EACH CLASSIFICATION.
51	to use sarcasm or some funny words just to hurt. The third	one is to dig at someone’s various faults by saying bad things
52	Cessation even if the Truth of the Path is practiced. The third	one means not believing in the Three Jewels and slandering them. b) Three
53	competitor out of the fear that he will best you. The third	one means the desire to kill and so forth while holding past harm
54	means to mature the Dharma qualities of your mind; and the third	one means to fully mature sentient beings. IV. CHARACTERISTICS OF EACH CLASSIFICA
55	evil deeds is fearsome, so one should feel remorse. (c) The Third	One, Necessity of Being Free of Nonvirtue Quickly. You may think that it
56	ld permanence and solidity through not understanding cause and result The third	one - one is established in equipoise and when one realizes all phenomena as
57	second one extends from the first to the seventh bhumi. The third	one ranges from the eighth to the tenth bhumi. The fourth one is
58	the arising of hatred, out of resentment, or in competition. The third	one refers to making sacrifices and so forth. b) Three Results of Taking
59	breaking into a house without others noticing and so forth. The third	one refers to deceit through measurements, scales, and so forth. b) Three Results
60	a difference between harm and benefit for oneself and another. The third	one refers to a lie with no benefit or harm. b) Three Results
61	second one means to divide two friends with indirect language. The third	one refers to secretly dividing. b) Three Results of Divisive Speech. “Result of
62	all the worlds, And afflicting emotions will not increase. c) The Third	One. The principle objective is to increase meditative concentration quickly. The s
63	. These are to be known by the [first] three examples, the [fourth]	one, and the [following] five. The dharmakaya is to be known [in] two
64	. These are to be known by the [first] three examples, the [fourth]	one, and the [following] five. One may wonder: What are the three aspects
65	third one ranges from the eighth to the tenth bhumi. The fourth	one is the level of Buddhahood. Thus, the Ornament of Mahayana Sutra says:
66	one is the method for increasing the strength of bodhicitta. The fourth	one is the method for deepening bodhicitta. The fifth one is the method
67	bodhicitta. The fourth one is the method for deepening bodhicitta. The fifth	one is the method for not forgetting bodhicitta. 1. Not Forsaking Sentient Beings f
68	that there is no other creator 4. Showing the connection with the sixth	one, the mind B1B1. Showing the dependent origination of the five engaging
69	deeds and good fortune. B 1B4. Showing the connection with the sixth	one, the mind Even the relation of the mind and dharmas Is like
70	the sound made by those experiencing the unbearable cold. In the sixth	one, the skin turns blue and cracks into five or six pieces like

71	and tangibles - and the group of inner cognitions, together with the seventh	one, is the alaya consciousness itself, which has the power to obscure the
72	cracks into ten or more pieces like lotus petals. In the eighth	one, the color turns to a darker red and the body cracks into

Appendix D: Generic pronoun *one* as the sentence object

Examples of the theme: cause and effect; positive, negative or neutral effect on *one*

#	concordance		effect
1	tes of the worlds of sentient beings. [8,64] The special support, [which makes	one] a suitable vessel for the sacred teachings, is the possession of all	neutr.
2	Realm. The reason for this is that the above meditations will bring	one an experience of the peace of perfect meditation and will free from	+
3	disease changes, one's mind changes, and the passage of time brings	one closer to death. First, "one's body changes" means that once your	-
4	you say or do, and are confused. "The passage of time brings	one closer to death" means that breath becomes short and labored, and you	-
5	vehicles— Therefore, it is the noble refuge. Thus, the common refuge protects	one from all harms, the three lower realms, unskilful means, and belief in	+
6	ght of the dhyanas and the formless. [12,66] The remaining eight can extricate	one from attachment to their own levels and to the levels above. [12.67] The	+
7	ree kayas. [13,19] Concerning the 'Dharma', the truth of the path, which frees	one from attachment, has the [three] sun-like qualities of 'purity' since it	+
8	his life. Further, it nourishes faith, supports perseverance, and quickly frees	one from attachment and hatred. It becomes a cause for the realization of	+
9	latter three constitute the cause, the truth of the path, which frees	one from attachment (8). Thus it has the characteristics of the two truths, which	+
10	for enlightenment. The second is contempt for "inferior beings," which hinders	one from [developing] love and compassion for others. The third is distorted percep	-
11	to abandon these: The first of these faults is faintheartedness, which hinders	one from exerting effort and striving for enlightenment. The second is contempt for	-
12	ence inspired by delight in enlightenment. There will be mindfulness preventing	one from forgetting the means to enlightenment, and meditative stability that is on	+
13	nd upasika. These are divided into laypersons and renounced. They all restrain	one from harming others. The pratimoksa vows provide restraint only for one's	+
14	the path of junction, one becomes free from the veils that prevent	one from seeing the two buddhakayas. Through the cultivation of this path one'	+
15	is supported by the 'capable [preparatory] stage', it is able to free	one from the attachment of the nine [levels] - those of the desire [realms],	+
16	he primordial wisdom of sameness, great wisdom, which is completely pure, frees	one from the faults of existence and suffering, and by great compassion one	+
17	aggregates. *1 The Ornament of Mahayana Sutra says: Perseverance will liberate	one from the view of the transitory aggregates. If one has perseverance, one	+
18	ned by abandonment B.II.2.2.2.1.2.2.2.2.1.3. The way this attainment liberates	one from the two extremes B.II.2.2.2.1.2.2.2.2.2. The function being what causes a	+
19	e, annotation 26.) B.II.2.2.2.1.2.2.2.2.1.3. The way this attainment liberates	one from the two extremes Their analytical wisdom has cut all selfcherishing witho	+
20	three reasons for the certainty of death: a. because there is no	one from the past who is alive, b. because this body is composite,	-

21	moment, death will definitely occur. a. My death is certain because no	one from the past is alive. Acharya Ashvaghosha said: Whether on the earth	-
22	, unskilful means, and belief in an abiding person. The special refuge protects	one from the lower vehicles and so forth. 8. Training. There are three general	+
23	does not contain any mental poisons but constitutes the obstacle that hinders	one from understanding the knowable. The third-mentioned veil is a particular aspec	-
24	is the attitude of gladly engaging in what is virtuous. It makes	one fully accomplish what is virtuous. [1,58] Within the non-virtuous mental states	+
25	path, the Dharma which becomes the path, and the Sangha which guides	one in order to accomplish the path. The Abhidharma says: What is faith?	+
26	is for the achievement of enlightenment, the contributory causes that encourage	one in practice, the method of practice, the result that is accomplished, and	+
27	either because of oneself or the Dharma. Its function is to support	one in refraining from negative actions. [1,52] Shame has the function of causing	+
28	several throw one into a single [rebirth], or that one action hurls	one into a single rebirth. In this way, here it should be understood	neutr.
29	a single impelling action throws one into several rebirths, that several throw	one into a single [rebirth], or that one action hurls one into a	neutr.
30	in that rebirth. It is possible that a single impelling action throws	one into several rebirths, that several throw one into a single [rebirth], or	neutr.
31	towards a sentient being, a painful object, or pain [itself]. It makes	one not abide in peace and creates the basis for negative action. [1,63] Arrogance	-
32	towards a sentient being or an object that causes pain. It makes	one not become involved in negative actions. [1,55] Non-delusion means being with	-
33	about the meaning of the [four] truths. Its function is to make	one not engage in what is virtuous. [1,65] Belief is the view of all	-
34	the absence of desire towards [samsaric] existence or worldly things. It makes	one not engage in negative actions. [1,54] Non-aggression is the absence of a	+
35	being without delusion concerning what is true due to discrimination. It makes	one not engage in evil deeds. [1,56] Non-violence is a compassionate attitude belo	+
36	this? Again, some say that the buddha qualities do not exist in	one primordially but they are newly established from the seeds of hearing and	neutr.
37	bound there, unable to attain the non-abiding nirvana. Furthermore, it binds	one there permanently according to the assertions of the three-vehicle system. Even	-
38	nt quickly. The Ornament of Mahayana Sutra also says: Perseverance will allow	one to achieve supreme enlightenment. The Sagaramati-Requested Sutra says: For	+
39	mental vow is a continuous intention, together with its seed, which forces	one to also obtain the vows of body and speech. The dhyana vow	+
40	as well as the knowledge of learning, reflection, and meditation which enables	one to attain this result. One then endeavors in the practices of cultivating	+
41	phical schools according to the inner science. Understanding of them will cause	one to be learned in the meaning of what is correct or incorrect. [5,14]	+
42	will also cause the discriminating mind to decrease. It will always cause	one to be far from primordial wisdom. And: One who is attached to	-
43	an intention to cause harm, and to refrain from forgiving. [1,77] Spite causes	one to be unforgiving and utter harsh words out of fury or resentment. [1,78]	-
44	harm to others and having harmful motives. The bodhisattva’s vow causes	one to benefit others. Without avoiding harm, there is no method of benefiting	+
45	by hitting them. [1,76] Resentment belongs to the category of anger. It causes	one to cling to an intention to cause harm, and to refrain from	-

46	ave much enmity, fear and distraction; the 'three wrongdoings' since they cause	one to engage one's three doors in perversion; the 'three losses' because	-
47	and so forth to feel weariness. They will eradicate conceit and induce	one to enter the path. [13,38] Various notions of questioning have been taught, su	+
48	ngs defilements everywhere there is purity and creates discrimination. It binds	one to existence and is the root of self-clinging. It is unawareness,	-
49	, but if the remaining three [bonds] are not relinquished, they will cause	one to fall back again. It is taught that one will become a	-
50	subsequent cognizance of the dharma of suffering is the wisdom that allows	one to grasp reality and realize it as it is. The cognitions such	+
51	truths; the 'three prospects of desire' and so forth because they cause	one to have much enmity, fear and distraction; the 'three wrongdoings' since they	-
52	not engage in virtue; instead, one creates nonvirtue which completely binds	one to misery [in subsequent lives]. [10,52] The same way, there are: ii-iv)	-
53	of stinginess'. They are called 'bonds' since each of these also binds	one to misery in subsequent lives due to engagement in nonvirtue. [10,53] These [b	-
54	, the lowest of the three realms. Pleasure-seeking and ill-will cause	one to never transcend the desire realms. One may have abandoned the desire	-
55	Sutra says: Since the practice of wisdom awareness without method will bind	one to nirvana, and the practice of method without wisdom awareness will bind	+
56	subsidiary disturbing emotions: [1,75] Fury is the increase of anger. It causes	one to prepare to harm others, such as by hitting them. [1,76] Resentment belongs	-
57	regret', and v) the ['veil of] doubt'. Taken in succession, they cause	one to refrain from virtue: at the time of wishing to take ordination,	-
58	MEDITATION. Awareness of the impermanence of all composite phenomena leads	one to release attachment to this life. Further, it nourishes faith, supports perse	+
59	In addition], they are called 'emotional obscuration' because they totally bind	one to samsara, whereby one fails to engage in virtue but carries out	-
60	to nirvana, and the practice of method without wisdom awareness will bind	one to samsara, it therefore becomes necessary to unify them. The Sutra Shown	-
61	in refraining from negative actions. [1,52] Shame has the function of causing	one to shun misdeeds, either because of being reproached by other [noble] people	-
62	and death for a long time; the 'three agonies' since they cause	one to spin [through samsara] and harbor doubts about the Three Jewels and	-
63	are completely tormented by them; the 'three jungle chains' because they force	one to take all kinds of rebirth in the jungle of existence; and	-
64	rigin Karma [9,1] These defiling samsaric aggregates, the causes that induce	one to take rebirth within samsara, do not originate without a cause nor	-
65	next existence. The seed placed in the [all-ground] consciousness which propels	one to the [next] rebirth is called impelling consciousness, and that which leads	neutr.
66	' because they corrupt one's discipline; the 'three evils' since they cause	one to undergo birth and death for a long time; the 'three agonies'	-
67	] mind and the mental states, the basis; the sustenance of will impels	one towards the following life; and the sustenance of consciousness actualizes that	neutr.
68	finger can be of benefit, Buddha said that even if it makes	one uncomfortable, Helpful things should be done. You should not give traps or	neutr.
69	appearance of the cause and effect of things. [7,4] These three times make	one understand that [a thing lasts] for such and such [a duration] by	+

Appendix E: The generic pronoun *one* as the subject of a sentence in conditional subordination

#	concordance	
1	pure, the purification of wrong-doing is needless for a practitioner. 3. If	one abides by the natural path53 the accumulation of gathering (Sambhar) is needles
2	, it is the morality associated with the Hearer's renunciation. c) If	one accepts them with an attitude of achieving the great enlightenment, it is
3	to the pratimoksa precepts, depending on one's mental state: a) If	one accepts these seven types merely from a desire to have the happiness
4	virtue to enlightenment, it will not be exhausted between now and when	one achieves the heart of enlightenment. VI. PERFECTION. Concerning the perfect p
5	not have moral ethics. Engaging in the Middle Way says: Even if	one achieves wealth through generosity, The being who breaks his leg of moral
6	. Relating to object, one will be born in the hell realm if	one acts nonvirtuously toward beings of higher status; if toward mediocre beings, o
7	of afflicting emotions, by the frequency, and by the object. First, if	one acts with hatred, one will be born in the hell realm. If
8	acts with hatred, one will be born in the hell realm. If	one acts with desire, one will be born as a hungry ghost. If
9	acts with desire, one will be born as a hungry ghost. If	one acts with ignorance, one will be born in the animal realm. The
10	rification of generosity, the Collection of Transcendent Instructions says: If	one acts with emptiness and the essence of compassion, All the merit will
11	Son Sutra says: Anger is not the path toward enlightenment. Therefore, if	one always meditates on loving-kindness, enlightenment will be produced. II. DEFI
12	the mind. Naropa said: Mind is the base of all phenomena. If	one analyses the nature of the mind by four reasonings, He will found
13	likes of Chya or Ishvara have composed various treatises explaining that if	one analyses the causes and conditions of amazing appearances – such as the brillia
14	, taste and touch. From what causes and conditions do these arise? If	one analyses them, one ascertains that the views of worldly people, tirthikas, Vaib
15	r defilements: Self-view, self-confusion, Self-pride and self-supremacy. If	one analyses them using the argument of the freedom from one [and many],
16	so on are just dream objects, only the appearance of mind. When	one analyses them through the method of the 'one and the many', one
17	is suitable for that. Just as dharmas appear in this way, if	one analyses with reason those things that are established to be or not
18	directions, such as east and so on, there are 1,560 self views. If	one applies self and other, there are 3,120 self views. Because it is the
19	as arising from the skandha - in total twenty self views. Furthermore, if	one applies the skandhas of past, present and future, there are sixty self
20	skandhas of past, present and future, there are sixty self views. If	one applies the skandhas of the twenty-six directions, such as east and
21	all sentient beings, limitless as space, to have happiness and benefit. When	one arouses this kind of mind, it is called genuine loving-kindness. The
22	a hole in a wall and so forth and go there. When	one arrives, if one is going to be reborn a male one develops

23	the dharmakaya, the appearance of trainees and previous aspiration prayers. If	one asks how buddhas are not swayed by dualistic discriminative understanding while
24	say it is provisional meaning 2E I. The necessity of explaining If	one asks if it is not even an object of the bodhisattvas who
25	ability to create and obscure. Therefore the manas has two aspects. If	one asks what, in summary, are the characteristics of this dual consciousness, they
26	II BI D. How there is ascertainment of three qualities Therefore if	one asks what is the reason for the buddha-nature being termed 'dharmakaya'
27	III C. How the buddha-phase will not revert to impurity If	one asks whether or not the buddha-nature, so named in all three
28	? In that way, since they are only one’s own appearances, when	one asks whether or not the grasping person with incorrect conceptualization is dec
29	. The position here in the vehicle of the Buddha is that if	one asks which came first out of the four coarse elements and the
30	, rejection. The one who has this acceptance or rejection is deluded. If	one asks why this is so, it is because the outer objects are
31	, lower rebirths And samsara. There are no such kinds of worry. If	one asks why, it is because there is no foundation for the self
32	cannot follow the direct meaning, the true nature of the Mahayana. If	one asks why this is so, the answer is as follows. This emptiness
33	he arising of this unattached and unobstructed freedom is buddhahood itself. If	one asks why, the answer is as follows. Since it arises through yogic
34	, water, fire, and so forth. “Power over aspiration prayers” means that if	one aspires to perfectly benefit oneself and others, it will be accomplished. “Pow
35	beings have been acting in this error since beginningless time, then if	one attained Buddhahood by depending on this error, all would have attained enlight
36	ispelling suffering and establishing the happiness of all sentient beings. When	one attains Buddhahood, there are no conceptual thoughts or efforts. Therefore, can
37	he great accumulation path, one can attend a Nirmanakaya spiritual master. When	one attains the bodhisattva’s level, one can attend a Sambhogakaya spiritual master
38	suffering. When can this confusion be transformed into primordial wisdom? When	one attains unsurpassable enlightenment. If you think that perhaps this confusion
39	arises and the two sticks themselves disappear in the fire. Similarly, if	one attends to what is to be abandoned through the antidote, then at
40	eventually change to suffering. The Letter to a Friend says: Even if	one became a universal monarch, One would fall into slavery in samsara. Not
41	initive meaning. Do not rely on consciousness; rely on primordial wisdom. If	one behaves in accord with this teaching, since one will enter into the
42	root of samsara, then won’t one be liberated from samsara if	one believes in nonexistence? This latter view is a greater fallacy than the
43	aspiration bodhicitta, it can be restored by taking the mind again. If	one broke the action bodhicitta vow through loss of the mind of aspiration,
44	aspiration, it is restored automatically by restoring aspiration bodhicitta. If	one broke the vow through other causes, it should be taken again. If
45	himself in front of an image of the Thus-gone One. If	one can find neither a spiritual master nor an image, then one should
46	false. So although emptiness is the antidote to grasping at existence, if	one clings to emptiness it is an incurable view. As it says in
47	are four types of offenses which cause bodhicitta to be lost if	one commits them through the heavy afflicting emotions. Mediocre and small evil off

48	be taken again. If through the four offenses, confession is sufficient when	one committed mediocre and small evil deeds. Twenty Precepts says: If one loses
49	a long time - many billions of years - instead of being instantaneous?' If	one compares this with the beginningless time of consciousness, a billion years is
50	warmth when it is cold are the activities of the buddhas. If	one considers this properly, since it stands perfectly to reason, one will see
51	By ignorance, one will be born an animal. Relating to frequency, when	one creates countless nonvirtuous actions, one will be born in the hell realm;
52	, a great result will ripen. The Verses Spoken Intentionally say: Even if	one creates small merit, It will lead to great happiness in the next
53	ievement of complete enlightenment is explained in the Bodhisattva Bhumis: When	one cultivates that mind, one will not remain in the two extremes *16 Quickly,
54	the Mahayana family one cannot receive the bodhisattva's vow even if	one cultivates the mind through ceremony. Therefore, all the necessary elements mus
55	the accumulation of gathering (Sambhar) is needless for a practitioner. 4. If	one cultivates the natural state of mind, it is needless for a practitioner
56	de toward all sentient beings. Explanation of the first unwholesome deed. When	one deceives the spiritual master, abbot, master, or one worthy of offerings by
57	will not be exhausted until the end of the kalpa. Likewise, when	one dedicates the root of virtue to enlightenment, it will not be exhausted
58	generosity infinite through the power of dedication. It increases infinitely if	one dedicates this generosity practice to the unsurpassable enlightenment for the b
59	silver, until the ore is smelted, the silver does not appear. If	one desires molten silver, one must smelt the ore. Likewise, all phenomena are
60	-interest in this way, It becomes supremely beneficial for oneself. But if	one develops loving-kindness and compassion, then one is attached to sentient being
61	Joy, is attained at the time of the path of insight when	one directly realizes the meaning of all-pervading emptiness. The second to tenth
62	of the body stand up, that is called great loving-kindness. If	one directs this kind of mind toward all sentient beings equally, it is
63	the mind, one enters into the Mahayana, the unsurpassed enlightenment. b) If	one does not have the desire to achieve Buddhahood, which is called the
64	, one will gain prosperity through the practice of giving wealth, even if	one does not wish it. Furthermore, one can gather trainees through generosity and
65	urpassable, perfect, complete enlightenment. In the conventional state, even if	one does not so desire, one will achieve the perfect happiness of samsara.
66	a universal chakra monarch in all one's different lifetimes, even if	one does not so desire. Engaging in the Conduct of Bodhisattvas says: While
67	the four immeasurable will acquire the attainment of Nirvana. Yet even if	one does not acquire this highest goal he will acquire a happiness which
68	very important to generate bodhichitta towards the beings of this life. If	one does not have it, one should take them as one's object
69	in my biography, "One may exert oneself for an aeon but if	one does not realise this vital point, nirvana remains unattainable". 2D III. Show
70	, it is called immeasurable loving-kindness. E. Measure of the Practice. When	one does not desire happiness for oneself, but only for other sentient beings,
71	a definitive meaning, and a definitive meaning a provisional one. For if	one does so, the way in which the Perfect Buddha [has expounded] his

72	an Buddha's definitive meaning, the object of meditation. Accordingly, even if	one engages only in hearing and thinking, there will be many beneficial consequence
73	called because their projection is space? *13 No. For the first three, when	one enters into the absorption, infinite space and so forth are projected in
74	is meditative stabilization. The Increase of Great Realization Sutra says: If	one enters such a meditative concentration for one moment, it has greater benefit
75	4,32] After having understood in general this way of a continual occurrence, if	one examines how many lives it takes to complete [one cycle of twelve
76	sional meaning dharma wheel. Therefore the ultimate stainless buddha exists. If	one examines the basis of these paths carefully it is buddha-nature itself
77	form and formlessness, which appear as the objects of deluded consciousness. If	one examines these opinions well with a straightforward mind, some have an incorrec
78	Sees their forms, fear arises. No need to mention what happens when	one experiences the intolerable result. Therefore, the result of evil deeds is fea
79	-four in all. So it is explained in the general way. If	one explained it more extensively, each would be found to have a million
80	, and the suffering of suffering would be like mold on fruit. *1 If	one explained these three sufferings with their definitions, the all-pervasive suff
81	on, as the absolute qualities that are by nature completely pure. When	one falls to assertion or denial, believing that the adventitious evils, which are
82	faults of samsara, One will develop a great sense of sadness. When	one fears the prison of the three realms, One will make an effort
83	forsake all sentient beings, neither will the hawk and wolf. Therefore, if	one forsakes even one being and does not apply the antidote within a
84	he created all fear and suffering, he could not be autonomous. If	one fully appreciates that there are such faults, one can understand that praising
85	, the self does not exist. The Precious Jewel Garland says: Therefore, when	one fully realizes as-it-is perfectly, the two do not arise. “Realizing
86	the human kingdoms are also a cause of suffering. Therefore, even if	one gained a universal monarchy over human beings, that would eventually change to
87	vessel well, Then you should give teachings even without a request. When	one gives a teaching, it should be at a clean and pleasing place.
88	to the gift. Also, one should not expect a result. Therefore, if	one gives everything with great skill There will be infinite virtue even if
89	of this meditation The Expression of the Realization of Chenrezig says: If	one had just one quality, it would be as if all the Buddhas’
90	continues, and there is love and kind support from others. Thus, when	one has all ten qualities, five from oneself and five from outside, they
91	[from the Ratnakuta Sutra] it says: In order to increase wisdom, If	one has both hearing and thinking, Then one should enter into practice. Through
92	intelligence, he remains as a speaker forever, which is a deviation. 2. If	one has deep faith and little intelligence, his effort goes meaningless, which is
93	also encompassed in this meaning. The Completely Non-Abiding Tantra says: If	one has eaten the food of uncontrived nature, one will satisfy all the
94	free such beings from these sufferings. E. Measure of the Practice. When	one has fully purified self-cherishing, is fully released or cut from the
95	or does not exist, is ignorant. One will not be liberated when	one has ignorance. E. Explanation of the Fifth, the Path that Leads to
96	please that people and follow his way. It is a deviation. 10. When	one has knowledge and spiritual power, but an unstable in mind and always

97	temples to the Tathagata Dawaytok Developed bodhicitta for the first time. If	one has little wealth, it is sufficient to make a small offering. In
98	, and exhort yourself to be diligent. The ten deviations from Dharma: 1. If	one has little faith and sharp intelligence, he remains as a speaker forever,
99	patience. When one has patience, one can make effort with perseverance. When	one has made effort with perseverance, meditative concentration will arise. When on
100	, one will accept the pure morality without focusing on material concerns. When	one has moral ethics, one will have patience. When one has patience, one
101	single blade of grass Developed bodhicitta for the first time. Again, if	one has no wealth, there is no need to feel sad over the
102	all sentient beings, and h) one quickly attains perfect enlightenment. a) If	one has not cultivated the supreme bodhicitta, he is not counted as part
103	of the Mahayana family even though he may have excellent behavior. If	one has not entered the Mahayana, one cannot achieve Buddhahood. But one who
104	long time and has practiced meditative concentration for millions of kalpas, if	one has not realized this perfect meaning, one will not be liberated, in
105	—is complete in this. The Vajra-like Meditative Absorption Sutra says: When	one has not moved from emptiness, the six perfections are included *17 The Brahma
106	material concerns. When one has moral ethics, one will have patience. When	one has patience, one can make effort with perseverance. When one has made
107	Perseverance will liberate one from the view of the transitory aggregates. If	one has perseverance, one will achieve unsurpassable enlightenment quickly. The Or
108	achieve the treasury of limitless primordial wisdom of the Victorious One. When	one has perseverance, one can cross the mountain of the view of the
109	path, mistake the path, and feel doubt about the [right] path, when	one has relinquished these three main [bonds], it has been taught, by implication,
110	into the unbearable hell As if it were a lotus lake. If	one has the attitude of wishing to accomplish the happiness and benefit of
111	velop the three moralities, *15 which constitute the bodhisattva training. When	one has the desire to attain Buddhahood, these three moralities are developed and
112	e primordial wisdom by oneself constitutes prostrations. It is also offering if	one has this. The Meeting of Father and Son Sutra says One who
113	is a deviation of whatever you do becoming a worldly decoration. 9. If	one has too much attachment and interest to one’s house, household articles
114	of bodhicitta are obtained through the accumulation of merit. So therefore, if	one has wealth it is not sufficient to offer just a little; one
115	unable to prevent them. In general, Buddhists say, 'It is good if	one has worldly pleasures and the causes of such pleasures, such as health,
116	by strengthening its family. “By the power of hearing” means that when	one hears different teachings, one also develops bodhicitta. “By the power of virtu
117	the buddha-nature mind the force of these qualities will manifest. If	one inquires what the reason is for their manifestation it is that, apart
118	ne has made effort with perseverance, meditative concentration will arise. When	one is absorbed in meditative concentration, one will perfectly realize the nature
119	, distress, and conflict. Therefore, no one is free from them. Thus, when	one is aware of the faults of samsara, one will withdraw from the
120	to be purified, the ground of purification; and 3. the purifier. Thus, if	one is being very concise, there are three points. As it says in
121	delicacy will not help, Then it should not be given. As when	one is bitten by a snake Cutting the finger can be of benefit,

122	beings with form [the links] from consciousness until becoming will occur. If	one is born in the Formless Realms, consciousness until becoming consisting of the
123	that virtuous deeds are the cause of higher rebirth or liberation. When	one is born at a time when the Buddha is absent, then there
124	one’s mind is the great offering which delights Buddha. Again, if	one is endowed with this, it is also the very purification of evil
125	pas of copying, reading, listening to, explaining, or reciting the Dharma. When	one is endowed with the meaning of emptiness, there is not a single
126	or stupid persons cannot understand the teachings on virtue and nonvirtue. When	one is free from all eight of these conditions, it is called the “
127	power over the pure fields. As it says in the Sutralamkara: When	one is freed from clinging, since pure enjoyment as desired is manifested, one
128	a wall and so forth and go there. When one arrives, if	one is going to be reborn a male one develops an attachment for
129	develops an attachment for the mother and aversion for the father. If	one is going to be reborn as a female, then one becomes attached
130	ceptual thoughts, accumulation of merit should not be discontinued. Thus, when	one is habituated this way, the equipoise and post-meditative state become undiffer
131	*3—will see the all-pervasive suffering as suffering, as, for example, when	one is nearly recovered from a plague and the small pain of an
132	aspirational actions. This is attained through persistent effort; otherwise if	one is not persistent, one cannot achieve this bhumi. During the second limitless
133	. These four types are related to an individual’s spiritual realizations. When	one is ordinary or just beginning, one cannot attend Buddhas and bodhisattvas who
134	subsequently feels regret, any virtuous or evil [action] will be exhausted; if	one is skilled in means, even a great misdeed can be quickly purified
135	people will not feel the all-Pervasive suffering as, for example, when	one is stricken with a serious plague and a small pain in the
136	in the intermediate state] resembles black cloth or pitch-black darkness if	one is to take birth in the lower realms, and resembles white cotton
137	in the lower realms, and resembles white cotton cloth or moonlight if	one is to take birth in the higher realms. [8,47] It is also said
138	beings in the lower realms and so forth The second one when	one is well trained in the practice of the Four Noble Truths, *4 understands
139	one of any of the three mental [nonvirtues] may be manifest. When	one kills out of ill-will, two are occurring together, and so forth.
140	in this than ordinary people who do not understand science, yet, if	one looks, one can directly see that they are troubled by sufferings. Although
141	evil offenses are merely disgraceful. Twenty Precepts says: In addition, when	one loses aspiration bodhicitta, it breaks action bodhicitta The Collection of Com
142	when one committed mediocre and small evil deeds. Twenty Precepts says: If	one loses the vow, one should take it again. If through the mediocre
143	do accept rebirth in that realm. [9,29] The dhyana vow is abandoned when	one loses the state of serenity. By shifting states, one obtains [the vow]
144	non-harmonious mind also breaks the vow. XI. METHOD OF REPAIRING. If	one lost aspiration bodhicitta, it can be restored by taking the mind again.
145	superior wisdom awareness. The Lamp for the Path to Enlightenment says: If	one maintains the vow of action bodhicitta And trains well in the three
146	: Realization will not arise from cast images and so forth. However, if	one makes energetic effort in bodhicitta, from that the yogin will become the

147	ue, generosity, and so forth. The Teaching Suchness Sutra says: Shariputra, if	one meditates on the meditative stabilization of suchness for even a single finger
148	the imprints of emptiness, the imprints of substantiality will be abandoned. If	one meditates that there is nothing whatsoever, in the future this will have
149	the one-vehicle system, one will be bound there for 84,000 kalpas. If	one only depends on method without wisdom awareness, one will not cross beyond
150	as a temple or an animal trap, ceases to exist, or when	one passes away. Virtuous [limited] vows are abandoned when cutting the root of
151	virtues, one will cherish this precious bodhicitta highly. In this way when	one practices, one sustains this mind without weakening. Therefore, one should pers
152	assable enlightenment. The 700 Stanza Perfection of Wisdom says: Manjushri, if	one practices the perfection of wisdom awareness, that great bodhisattva will quick
153	-fulfilling jewel has no conceptual thought, it manifests whatever one needs if	one prays to it. Likewise, depending on the Buddha accomplishes all the purposes
154	? Yes, very much so. The Sutra of the Great Parinirvana says: If	one purifies evil deeds through remorse and repairs them, they are purified as
155	and result The third one - one is established in equipoise and when	one realizes all phenomena as the nature of emptiness, compassion arises, especiall
156	cycle continues. b) The interdependence of nirvana is in reverse order. When	one realizes all phenomena as the nature of pervading emptiness, then ignorance cea
157	is like a hell. One has to endure inconceivable suffering. So, if	one realizes this, how can one think of entering into the mother’s
158	eficial effect of establishing all the favourable conditions. In this way when	one recollects all these virtues, one will cherish this precious bodhicitta highly.
159	the same result even though there is but a single doer. If	one rejoices in any virtuous or evil [action], one will achieve [a result]
160	turn into the substance of a buddha, a King of Munis, when	one relies on the necessary condition, which is the virtue of the two
161	is luminous self-cognition in the present moment. It is like when	one remembers a vase one saw earlier: in that experience, the thing which
162	, since one moment does not exist, the mind is also nonexistent. If	one says there are many moments: if a single moment existed, then it
163	happiness and suffering come? What is pleasure and what is pain? If	one seeks truth, Who is there to desire and what is there to
164	feel great fear. Nagarjuna said in the Letter to a Friend: When	one sees a drawing of hell realms, Or hears, recollects, reads about, or
165	to be added. This is the view of reality in reality. When	one sees reality, it is complete liberation. Jetsun Zhepa Dorje says: Regarding
166	pleasures of samsara. The Meeting of Father and Son Sutra says: When	one sees the faults of samsara, One will develop a great sense of
167	snow mountain during the day are experienced in one's dreams when	one sleeps, such that one actually feels it is daytime. The discrimination of
168	evil [action], one will achieve [a result] similar to the doer; if	one subsequently feels regret, any virtuous or evil [action] will be exhausted; if
169	free from defilement, possesses [seven] particular properties corresponding, if	one takes an example, to the particular properties of a lamp. These are
170	to understand that buddhahood possesses the quality of being uncreated. Yet if	one takes it as a whole as being uncreated, one needs to understand
171	three realms, then this is morality with a vested interest. b) If	one takes these precepts in order to completely free oneself from all suffering,

172	takes an ordinary person along on the ship of the conquerors. If	one tastes the drop of nectar of this teaching, Given to liberate minds
173	union of appearance and emptiness – is like the moon in water. If	one then properly inquires as to the source, or cause, from which these
174	about And do not think about what cannot be thought about. When	one thinks about neither the thinkable nor the unthinkable, Emptiness will be seen
175	to subdue attachment to self would also be of provisional significance. If	one thinks it is contradictory to the true nature of all dharmas being
176	are dependent on the perceiver and, in that way, are undeceiving. [13,12] When	one understands in this manner, one becomes undeluded about the meaning of the
177	[Root Verses of the Wisdom of the] Middle [Way] it says: If	one understands that dharmas are empty, The dependence of actions and effects Is
178	abandon an attitude attached to emptiness. In the Sutralamkara it says: When	one understands that there is nothing other than mind, one realizes that mind
179	lack independent control, and consequently there would be no benefit even if	one were to worship it. In these and other ways, the concept of
180	duration it takes to pierce each is therefore understood through inference. If	one were to calculate with the sometimes mentioned sixty-four or five [leaves]
181	to arise in one's [stream-of-being] within this lifetime. If	one were to die without having destroyed the [evil deed] with a remedy,
182	: suffering of the lower realms and suffering of the higher realms. If	one were to explain the first type, the lower realms consist of these
183	state of buddhahood. Just as one could not see the sun if	one were to eliminate its light and its rays. In this way the
184	state of buddhahood. Just as one could not see the sun if	one were to eliminate its light and its rays. Therefore, as has been
185	one could not see the completely pure orb of the sun if	one were to eliminate its clear light and its beaming rays. B.II.2.2.2.1.2.3.
186	based on necessity in regard to the inclinations of the disciples. [10,100] If	one wishes to make the above divisions based on the system of the
187	shall I too Successively follow the practices. Repeat this three times. If	one wishes to cultivate bodhicitta and take the vow separately, then recite the
188	benefit of all sentient beings. As soon as it is attained, if	one wishes to, one has the ability to: (i) display one hundred bodies,
189	jewels. This shastra Arose from the heart of a supreme arya. If	one wishes to enter the castle of the Mahayana teachings, Like a magical
190	about neither the thinkable nor the unthinkable, Emptiness will be seen. If	one wonders how emptiness can be seen, it is said in the Accomplishment
191	, the wise one does not abide In either existence or nonexistence. If	one wonders what the middle is, which avoids the two extremes, the Heap

Appendix F: Collocates of the node *bodhicitta* (MI and LL) in MDSTB corpus

5-5 window span; min frequency of 3

Rank	frequency	frequency (l)	frequency (r)	stat	collocate
1	5	2	3	10.29951	objectives
2	3	3	0	9.34015	engendered
3	8	1	7	9.24061	lost
4	3	2	1	9.1475	recollecting
5	12	11	1	9.01822	cultivated
6	31	26	5	9.00894	aspiration
7	14	11	3	8.94363	cultivating
8	28	26	2	8.87804	cultivation
9	8	6	2	8.8621	developed
10	15	15	0	8.79701	cultivate
11	13	9	4	8.64827	relative
12	34	27	7	8.50026	action
13	12	9	3	8.31461	beneficial
14	7	4	3	8.28243	ceremony
15	6	6	0	8.24061	rise
16	5	2	3	8.12958	preparation
17	11	9	2	8.06778	effects
18	3	3	0	8.06004	generated
19	32	16	16	8.02338	bodhicitta
20	4	3	1	7.70456	receive
21	11	3	8	7.70004	vow
22	3	2	1	7.62394	losing
23	4	4	0	7.60834	generate
24	15	11	4	7.48075	training
25	6	1	5	7.44706	ever
26	3	3	0	7.44706	classifications
27	6	0	6	7.36614	lamp
28	9	8	1	7.27303	method
29	4	2	2	7.24061	topics
30	6	1	5	7.24061	antidote
31	3	2	1	7.19331	motivation
32	4	3	1	7.03898	comprise
33	3	3	0	6.93805	primary
34	6	3	3	6.93805	holding
35	4	1	3	6.89011	toward
36	3	1	2	6.75519	brief
37	5	1	4	6.73472	instructions

Rank	frequency	frequency (l)	frequency (r)	stat	collocate
38	7	2	5	6.65565	ornament
39	3	0	3	6.65565	obtained
40	4	3	1	6.6318	develop
41	4	3	1	6.60834	practicing
42	5	1	4	6.59907	vows
43	3	1	2	6.53279	requested
44	15	12	3	6.52106	ultimate
45	5	2	3	6.18403	special
46	16	6	10	5.97008	first
47	7	5	2	5.83074	order
48	6	3	3	5.78995	mahayana
49	8	3	5	5.74876	types
50	4	4	0	5.74236	chapter
51	4	3	1	5.6799	tathagata
52	4	4	0	5.66772	characteristics
53	3	1	2	5.63971	realize
54	4	3	1	5.6318	given
55	4	2	2	5.59676	after
56	10	5	5	5.43738	sutra
57	7	4	3	5.42153	supreme
58	3	1	2	5.41958	times
59	4	1	3	5.40267	some
60	4	0	4	5.40267	jewel
61	3	1	2	5.39261	make
62	3	3	0	5.31461	give
63	3	0	3	5.11408	wish
64	12	3	9	5.06535	if
65	3	2	1	5.06004	arises
66	3	2	1	5.01822	your
67	5	4	1	5.0141	teachings
68	4	1	3	5.0003	complete
69	35	17	18	4.92861	for
70	3	1	2	4.91868	second
71	4	3	1	4.90433	person
72	4	1	3	4.89011	born
73	18	11	7	4.87604	has
74	4	3	1	4.86905	liberation
75	8	6	2	4.83462	having
76	8	5	3	4.82107	enlightenment
77	12	6	6	4.75949	two
78	18	9	9	4.75233	mind
79	3	1	2	4.73811	thought
80	5	1	4	4.73472	present

Rank	frequency	frequency (l)	frequency (r)	stat	collocate
81	7	3	4	4.70812	what
82	4	1	3	4.65565	essence
83	5	3	2	4.64606	bodhisattva
84	4	4	0	4.60254	bodhisattvas
85	6	1	5	4.59676	time
86	4	1	3	4.57955	ten
87	10	3	7	4.50076	through
88	9	3	6	4.48929	sentient
89	3	1	2	4.48217	terms
90	7	6	1	4.47205	cause
91	3	2	1	4.44014	realization
92	4	2	2	4.34337	power
93	3	1	2	4.23461	buddhas
94	4	1	3	4.21381	practice
95	59	25	34	4.19637	in
96	11	9	2	4.10311	who
97	8	4	4	4.07872	when
98	16	8	8	4.06868	with
99	3	2	1	4.06535	causes
100	8	6	2	4.05871	says
101	4	0	4	4.04684	said
102	3	2	1	4.04421	then
103	24	11	13	4.01822	one
104	4	2	2	3.96263	therefore
105	11	5	6	3.92657	will
106	4	3	1	3.90076	does
107	4	1	3	3.84144	at
108	21	11	10	3.83397	not
109	11	8	3	3.80523	on
110	9	3	6	3.77536	three
111	4	2	2	3.75519	them
112	48	32	16	3.75492	to
113	3	2	1	3.74662	can
114	3	0	3	3.74236	perfect
115	5	3	2	3.73726	four
116	132	88	44	3.73147	of
117	225	85	140	3.71382	the
118	11	6	5	3.70569	which
119	6	0	6	3.69423	s
120	15	4	11	3.67908	be
121	5	0	5	3.67745	path
122	19	11	8	3.6722	this
123	4	1	3	3.52912	thus

Rank	frequency	frequency (l)	frequency (r)	stat	collocate
124	57	11	46	3.52218	is
125	4	0	4	3.51815	should
126	17	6	11	3.48824	by
127	5	2	3	3.48786	i
128	9	3	6	3.48335	these
129	9	4	5	3.41958	b
130	9	5	4	3.41732	or
131	4	0	4	3.35553	he
132	6	5	1	3.3466	have
133	9	5	4	3.32839	they
134	3	1	2	3.32095	other
135	4	1	3	3.28409	great
136	4	2	2	3.23387	qualities
137	22	11	11	3.199	a
138	57	24	33	3.19852	and
139	7	5	2	3.15073	beings
140	3	1	2	3.14188	also
141	12	2	10	3.13278	from
142	7	3	4	3.08449	there
143	4	3	1	3.06269	an
144	6	3	3	3.03637	ii
145	3	1	2	3.00795	forth