PHRASAL VERBS IN UNDERGRADUATE WRITING: A FOCUS ON SOUTH AFRICAN SECOND-LANGUAGE STUDENTS

by

SUSAN IMMELMAN

submitted in accordance with the requirements for the degree of

DOCTOR OF PHILOSOPHY

in the subject

LINGUISTICS

at the

UNIVERSITY OF PRETORIA

SUPERVISOR: DR P.A. COOPER

OCTOBER 2024

Abstract

Global research into phrasal verb use suggests that second-language students struggle to attain competence in its use although it is prevalent in English and so key to English proficiency. The phrasal verb is a multi-part verb consisting of a verb proper and particle. This research study set out to discover whether this trend was also observable among South African second-language students, and, furthermore, to examine the general patterns of phrasal verb use by South African students, including first-language students. Two corpora, SAMuLCAT (comprising student writing samples from North-West University and the University of Pretoria) and WITS (comprising student writing samples from the University of the Witwatersrand) were used in the study, separated into first- and second-language subcorpora.

Phrasal verb use was approached from various angles to gain a detailed perspective on this grammatical feature. Investigation centred around phrasal verb occurrence and distribution, frequently used phrasal verbs, preference for phrasal verbs or their one-word synonyms, and phrasal verb error occurrence.

The results showed that second-language students use phrasal verbs more than first-language students, and that first-language students prefer one-word synonyms more than second-language students. Many high-frequency phrasal verbs were used by both groups of students. Phrasal verb distribution showed that a small percentage of phrasal verbs was used often, while the rest were used once only. More unnatural PV use was found in second-language than in first-language student writing which suggests that L2 students in particular would benefit from workshops on the appropriate use of phrasal verbs and their one-word alternatives.

Keywords

Phrasal verbs, one-word alternative verbs, student academic writing, second-language students, corpus linguistics, formulaic sequences, academic register, avoidance.

Acknowledgements

Great appreciation is due to my family for their patience and support while I was busy with this research, especially during the last few months when intense effort was required.

Prof. Lilli Pretorius and Dr Sally Hunt generously took the time to read sections of the research and offer suggestions for improvement. Their input was invaluable and is highly appreciated.

This research made use of two corpora. I am deeply grateful to Dr Trish Cooper, who gave permission for the use of the WITS corpus, and to Prof. Tobie van Dyk, who gave permission for the use of SAMuLCAT corpus, for making these excellent resources available to me.

The person to whom I owe the most, without doubt, is my supervisor, Dr Trish Cooper. I would like to thank her for her guidance, support and patience throughout this research study. The mark of a successful partnership is that it culminates in friendship.

Declaration

I declare that PHRASAL VERBS IN UNDERGRADUATE WRITING: A FOCUS ON SOUTH AFRICAN SECOND-LANGUAGE STUDENTS is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

28 October 2024

_____ _____

S. Immelman (Mrs) Date

Student no.: 04181956

Glossary of corpus acronyms

Terminology	Description
BNC	British National Corpus
BNC-AC	Academic writing subcorpus of the British National Corpus
CANCODE	Cambridge and Nottingham Corpus of Discourse in English
COCA	Corpus of Contemporary American English
COCA-AC	Academic writing subcorpus of the Corpus of Contemporary American English
GS-UK	Argumentative essays of British school leavers in the General Studies corpus
ICLE	International Corpus of Learner English
LOCNESS	Louvain Corpus of Native English Essays
LOCNESS-US	Argumentative essays of American L1 university students in the Louvain Corpus of Native English Essays

Table of contents

List of tables

List of figures

List of appendices

Chapter 1: Introduction

1.1 Introduction

Given the fact that we are a country of predominantly second-language speakers of English, English proficiency will remain in the forefront of the many educational issues with which we are faced, especially at tertiary academic level. Higher education in South Africa continues to show the effects of low academic literacy in English, as evidenced by high dropout rates among students whose mother tongue is not English (Carstens, 2016:203; Van Rooy & Coetzee-Van Rooy, 2015:31). Any research into English proficiency is, therefore, of value as it enhances the pool of knowledge with which various problems can be addressed.

The current research claims to contribute to this understanding of English within the academic sphere as it focuses on an area of the English language which appears to have attracted little attention in the South African context to date, namely the phrasal verb. The phrasal verb, examples of which include *make up*, *carry out* and *turn on*, falls within the category of formulaic sequences, and is one of the many multi-word structures in English which Gardner and Davies (2007:339) describe as "crucial to English" as it adds "a definite richness to the language". Furthermore, phrasal verbs are a regular feature of the English language (Chen, 2013a:420). For example, learners are likely to encounter a phrasal verb "in every 150 words of English they are exposed to" (Gardner & Davies, 2007:347). Phrasal verbs make up one third of all English verbs, and can thus be employed as one of the measures used to assess the level of English language proficiency (Riguel, 2014:111-112). Yasuda (2010:250) calls them "the most frequently occurring idiomatic strings of language in both spoken and written English". Hence, it can be said that:

> [f]or non-native speakers of English to sound like native English speakers, they have to develop their overall language skills and use phrasal verbs in their daily communication. This will simply open up a whole new world of mastering English language and vocabulary, and allow them to elevate their level to as close to native as possible (Haider, Saed, Hussein, Al-Abbas & Meqdadi, 2020:1185).

This study will investigate the use of the phrasal verb (PV) by native English speakers and second-language speakers at various educational institutions in South Africa. Should the findings of the investigation prove suitably informative, they may be regarded as contributing to the advancement of education in South Africa in general and English competence in particular.

In the next section, some background will be provided for the language issues faced by second-language students in an educational environment by means of a discussion on second-language acquisition in global terms, and also as seen from a South African perspective.

1.2 Background to the study

English is at present the medium through which most higher education is attained in South Africa. Consequently, it is essential that the many students whose mother tongue is not English are supported in their educational endeavours. This section first looks at the general issues underlying second-language acquisition, after which the issues pertaining to the South African environment in particular are discussed.

1.2.1 Overview of second-language acquisition

It is a phenomenon of the globalised world that most people speak two languages and that the second language most people speak is English (Macaro, 2010:3). Trebits (2009:472) describes English as "the most important language of communication in international and European business organizations […] and within EU institutions". In fact, the term *Euro-English* is now used to describe English as the *lingua franca* in Europe, since it is used "as a means of communication between native speakers of other languages" (Trebits, 2009:272).

Saville-Troike and Barto (2017:8-9) estimate that approximately 1.75 billion people worldwide are at present learning English as a second language, that there are more multilingual learners than monolingual learners, and that more learners are being educated in a second language than in their first language. Globalisation has, furthermore, encouraged the implementation of English as the *lingua*

franca in many universities around the world, necessitating proficiency in English in order to obtain academic success (Mathew, Nesi & Vincent, 2019). An example is given by Civan and Coskun (2016:2000), who, after investigating the effect of instruction in English at a Turkish university, found that "if students know English well, their academic success is positively affected from instruction in English". They further concluded that improving L2 students' English proficiency was likely to improve their "wellbeing during their education and after their graduation" (Civan & Coskun, 2016:2000).

It is worth noting that, while the term "second-language" learning is used, students might, in fact, be learning a third or fourth language (Saville-Troike & Barto, 2017:2) as is so often the case in the South African environment. Macaro (2010:4) makes the thought-provoking observation that "[a]s an object of learning, a second language has no rival", as "[n]o other subject is learnt by so many people over such long periods of time". He views second-language learning as, to some degree, a "messy, non-linear, but semi-permanent process" (Macaro, 2010:4). For this reason, Cook (2010:154) is possibly not entirely wrong in questioning the tradition of measuring L2 success against L1 competence:

> The child is learning their first language and has no other: the L2 learner is learning a second language when they already have a first. Why should the mastery of a second language be measured against the mastery of a first? L2 learners are only failures if they are measured against something they are not and never can be – monolingual native speakers.

It is useful to clarify what will be understood as a "second language" in this study. Saville-Troike and Barto (2017:4) give the following useful definition: "A second language is typically an official or societally dominant language needed for education, employment, and other basic purposes." Therefore, second-language learners are not necessarily learners of English, as, for example, in the case of an English speaker working in China, needing to acquire Mandarin to communicate effectively. It is thus important to clarify that second-language acquisition does not specifically refer to acquiring English, and that the discussion that follows is applicable to the acquisition of any second (or additional) language. In research, the second language is normally designated as L2, to distinguish it

from the student's first language (also known as the native language, primary language, or mother tongue), which is designated as L1 (Saville-Troike & Barto, 2017:4). Please note that the discussion in this section will cover global theories of second-language acquisition, and that the South African context will be the focus of the next section.

Second language acquisition has interested researchers for some time. While some form of second language learning can be said to have existed almost as long as language has been around, it was only in the 1960s that "the process of systematic reflection on language learning" emerged (Macaro, 2010:6), eliciting learning theories from linguistic, psychological and social disciplines, among others (Saville-Troike & Barto, 2017:26). Because of the involvement of such a mix of disciplines, with their differing theories and research methods, there is often disagreement as to which processes are the most effective (Saville-Troike & Barto, 2017:2). Nevertheless, the resulting interdisciplinary research has enriched this field of study, and continues to do so (Hummel, 2014:2; Saville-Troike & Barto, 2017:2; Spada & Lightbown, 2019). What the many theories of language acquisition that currently exist (Universal Grammar, Monitor Theory, various psychological perspectives such as Behaviourism, Cognitive Psychology, Connectionism, Processability Theory, and Vygotsky's sociocultural perspectives) (Spada & Lightbown, 2019) underscore is the fact that the acquisition of a second language is not without complications.

The 1980s saw an increased interest in the field of second-language learning because of a global surge in the need for learning English. Around this time, the idea that L2 learning was simply an extension of L1 learning was gradually superseded by an understanding that L2 learning required a completely new perspective and methodology (Macaro, 2010:6).

It is useful to have a brief overview of what L1 acquisition entails, as, according to Hummel (2014:6), L2 acquisition can only be understood in the light of first-language acquisition. There are currently two opposing views of what the process of L1 acquisition entails, namely the nativist approach (Saville-Troike & Barto, 2017) versus the emergentist approach (Van Rooy & Kruger, 2015). The nativist

perspective is that children are born with an innate language ability, which suggests that they do not learn their L1 necessarily by means of the imitation of adult speech (Saville-Troike & Barto, 2017:23). Rather, the process appears to take place spontaneously and subconsciously, with more structured teaching providing no further advantage. Moreover, children appear to absorb universal language rules instinctively over time, without these rules having been taught outright. Variation among L1 learners only surfaces in the rate of progress of language development, and is dependent on individual factors, such as ability (Saville-Troike & Barto, 2017:15).

It appears that L1 language development slows perceptively over time, so that, if any interference had been experienced during childhood, it is unlikely that complete fluency will be acquired (Saville-Troike & Barto, 2017:8). (There is a degree of dispute about when this change takes place, with estimates ranging from puberty to as early as age six (Hummel, 2014:23; Murphy, 2010:161).) That the ease with which the L1 language is learnt slows relatively early in life highlights the difficulties that L2 learners face when confronted by a new language in their adult years.

On the other hand, the emergentist approach denies the idea of an innate "prior design" that guides language acquisition (Van Rooy & Kruger, 2015:56), and proposes that "the exposure to actual language use in interactive contexts plays a central role" (Van Rooy & Kruger, 2015:42). According to this approach, language is acquired by means of constructions that are found to be useful for communication by the individual language learner, in conjunction with the possibility of cognitive processing by the learner. Meaning arises through a combination of form and function, yet cannot always be predicted accurately from its parts. Over time, as the language learner becomes familiar with a certain combination, its grammatical structure is inferred, its complexity disappears, and its meaning moves from the general to the abstract. Thus, "[w]hat is stored in the mind of the user … is a network of related linguistic constructions that emerge from interaction between people and exposure to data" (Van Rooy & Kruger, 2015:62). In the emergentist approach, the similarities rather than the differences between L1 and L2 acquisition are considered, in contrast to the nativist view that

focuses on the differences between the two. The rest of the discussion on second language acquisition will follow the nativist approach.

When the L2 is subsequently learnt, the acquisition process inevitably maps onto that of the L1 (Hummel, 2014:21), resulting in a process called "transference". The more similar the L2 is to the L1, the more likely that transfer will take place. The transfer from L1 to L2 can produce correct knowledge of the L2 (positive transfer), although the converse has also proved to be true: too much similarity can make the L2 learner suspicious about the validity of the transfer. On the other hand, many aspects of the L1 may not be relevant to the L2, or may be completely different to the L2 (negative transfer) (Saville-Troike & Barto, 2017:18). Negative transfer is sometimes referred to as "cross-linguistic influence" (Laufer & Eiliasson, 1993:36), or "interference", and is deemed "a major source of errors in the L2 learner's developing system" (Hummel, 2014:139). According to Hummel (2014:39), "[t]ransfer between languages is therefore subject to multiple influences and not fully predictable".

Besides transference, other processes might similarly have an impact on L2 learning. One such process is "overgeneralisation", where a correctly learned rule is applied inappropriately (Hummel, 2014:140). An example is L2 learners' habit of forming a plural for words which do not take a plural, such as "homeworks". This error is not a result of transfer from the L1, nor could it have emanated from imitating an L1 speaker, which means that the rule concerning plural formation had been internalised and was being put into practice. Thus, overgeneralisation is not an entirely negative phenomenon of second-language acquisition, as it demonstrates an attempt at productive language use.

Not all processes that are observable in L2 acquisition necessarily lead to errors. Formulaic sequences, for example, provide "building blocks" for L2 learners, in the form of "single unanalyzed units, or ready-made chunks" (Hummel, 2014:142). This allows learners to use phrases despite not knowing the meanings of the individual units of which such phrases are comprised. They are thus able to produce elements of the L2 that are beyond their current competence. Since they conform to this description of lexical units, PVs could be seen as such formulaic sequences.

Avoidance is a strategy that is used by L2 learners when they are faced with an unknown or difficult element in the L2 (Hummel, 2014:144). According to Hummel (2014:144), an undue emphasis by educators on detecting incorrect language use could obscure this problematic phenomenon. For instance, an L2 learner could, by using synonyms or rephrasing expressions, produce near flawless writing, while avoiding quintessential English language grammatical structures. When educators pay attention only to what is being produced by L2 students and ignore what is being avoided, they might miss the opportunity of genuinely advancing the L2 learner's English competence. Use of the avoidance technique has been noticed and reported in PV research. (See Chapter 2 for a further discussion of this phenomenon.)

Finally, acquiring the vocabulary of an L2 presents a singular challenge, yet its acquisition is crucial. There is no doubt, as Hummel (2014:147) contends, that "words are vital to the communication process". The challenges with vocabulary attainment for the L2 learner encompass the vast number of words that make up the L2 vocabulary (to which more are being added on a regular basis). Moreover, the L1 speaker presents words in a "continuous stream", which is overwhelming to the L2 learner until the skill of segmenting L2 communication into meaningful units is learnt (Hummel, 2014:148). To fully acquire a new word, several layers of knowledge about the word are required, among which are its meaning, its grammatical function, and its register (Hummel, 2014:148). These aspects are not obtained simultaneously, but rather with time and use, so that a word might be familiar to the L2 learner long before its correct use in a sentence is understood, or the correct register in which it should be used. A word that presents a particularly "complex grammatical and morphological structure" might also result in delayed acquisition (Hummel, 2014:149). Hummel (2014:149) further mentions semantic features that could prevent the assimilation of a word, such as its having multiple and metaphorical meanings. From these limitations, it seems inevitable that PV acquisition would cause problems for the L2 learner, given that PVs display complex multiword, separable structures, multiple, often idiomatic meanings, and a frequently informal register.

L2 acquisition appears to demonstrate some of the innate language learning strategies associated with L1 learning, although not to the same extent (Saville-Troike & Barto, 2017:23). For example, immersion in the new language is as necessary for the L2 learner as it is for the L1 learner, although the extent to which this is possible will necessarily vary (Hummel, 2014:23). Furthermore, repetition of words and concepts plays an important part in both L1 and L2 language attainment, as does the use of formulaic phrases (Hummel, 2014:23). Then, it is also true that both L1 and L2 learners are able to understand far more than they are able to produce (Hummel, 2014:23).

Nevertheless, given the differing stages at which the language is acquired, the L2 learner inevitably differs from the L1 learner. The most obvious difference is that the L2 learner is older, and, as Cook (2010:137) states, "age inevitably brings with it a host of factors that have little to do with language acquisition". There are also further differences, such as being more cognitively developed, with more experience that can be brought to bear on the process of acquiring the L2 (Cook, 2010:149; Hummel, 2014:21). (Hummel, 2014:21) also points out the emotional issues that might accompany the learning of a second language, such as feelings of alienation from the L1, anxiety about communicating in the L2, and pressure to attain proficiency in the L2. Added to this is the difference in learning environment from the L1, which normally takes place in the safety of the home or an informal setting, whereas the L2 is most often learnt in a classroom, which might be experienced as threatening (Hummel, 2014:21).

The L1 is also generally learnt in an immersive environment, while the L2 will probably be taught, at most, for a few hours a week (Hummel, 2014:21). It should be noted that the presence of technology "in all aspects of the lives of language learners" is changing this reality for many L2 learners (Chapelle, 2007). However, it must also be recognised that access to technology is not always available to many of the lower income groups in South Africa (Munje & Jita, 2020), with the result that the L2 students do not necessarily gain the advantages of immersion in English through online sources. For example, Munje and Jita (2020:274) observed a "lack of [Information and Communication Technology] resources in classrooms in selected South African Primary schools".

Because of the relative ease with which the L1 is generally acquired, many educational theorists insist on the L2 being taught from an early age. On the other hand, some research suggests that initially concentrating on supporting L1 development results in successful L2 acquisition later (Murphy, 2010:166). No definitive consensus has been reached in this regard, however. One solution, suggested by Murphy (2010:176), is informal L2 immersion from an early age, to create familiarity with the L2 before subsequent teaching takes place (much as was observed for L1 acquisition).

Researchers, furthermore, make a distinction between learning a second language in a formal setting, such as a classroom (what Cook (2010:151) calls an "artificial setting"), and an informal setting (what Cook (2010:151) calls a "natural setting"). In the South African situation, it is likely that L2 learning first takes place in a natural setting to some degree (the degree being dependent on the child's environment), followed by an artificial setting, the classroom.

It should also be noted that research into SLA is often conducted in, or by, "Western, educated, industrialized, rich and democratic (WEIRD) societies" (Henrich, Heine & Norenzayan, 2010), and that, what might be applicable to such societies are not necessarily applicable elsewhere. For example, Southwood, White, Brookes, Pascoe, Ndhambi, Yalala, Mahura, Mössmer, Oosthuizen and Brink (2021:14) suggest that, in the South African context, "a complex interplay of sociocultural influences on language development of young South African children may be present".

In this section, second-language learning was discussed as background to the difficulties faced by L2 learners. The particular issues faced by South African L2 learners will next be explored.

1.2.2 The South African context

While facing the challenges of second-language acquisition discussed in the previous section, South African L2 students also encounter difficulties that are unique to their environment. These difficulties are often, and rightly so, linked to pre-democracy policies that continue to leave their mark on the South African educational situation, as evidenced by the differences in throughput rates among

different groups of students (Van Rooy & Coetzee-Van Rooy, 2015:31). Yet educational difficulties also refer to current circumstances, as will be illustrated below.

Even though South Africa has eleven official languages, the general language of communication is English. When one takes into consideration that the most prominent mother tongue languages in South Africa are IsiZulu (24.4% of the population), IsiXhosa (16.3% of the population), and Afrikaans (10.6% of the population), and that English is the mother tongue of only 8.7% of the population, as at the time of the 2022 census (*Statistics South Africa*, 2024), it is clear that the majority of South Africans are L2 speakers. Van Rooy and Coetzee-Van Rooy (2015:32) point out the many problems that prevent L2 learners from attaining academic success at university, one of the primary issues being that of language, and learning in the L2 language, in particular. They further remark that "there is an underlying assumption that poor English language proficiency is a barrier to educational achievement" (Van Rooy & Coetzee-Van Rooy, 2015:32). L2 students are not only confronted by the overwhelming experience of being at a university, along with a sharp increase in cognitive expectation, but also by the fact that they must face these challenges in a language that is not their mother tongue (Carstens, 2016:204). This situation is likely to remain an integral part of the South African educational landscape for some time. The study of L2 learning in South Africa is consequently essential to the development of language skills, particularly within an educational context.

The profile of the L2 student in South Africa can, therefore, be said to differ somewhat from that of the L2 students described in international research. Whereas the L2 students described in such research, for the most part, refer to students with the same L1 attempting to learn English as a foreign language (that is, a language that is not the official language of the home country, nor is commonly used in that country) (Chen, 2013a; Dagut & Laufer, 1985; Hulstijn & Marchena, 1989; Laufer & Eliasson, 1993; Liao & Fukuya, 2004), the South African L2 students are not trying to learn a foreign language, but the *lingua franca* of their country. English takes up an interesting position in the South African context, in that it is the native tongue of only a minority of South Africans, with the result that

its use is regarded by many as oppressive because of its global hegemony (*Oxford English Dictionary*, 2023). Yet, for practical and economic reasons, it is the predominant language of communication, while "many black parents see it as a crucial instrument for their children's advancement" (*Oxford English Dictionary*, 2023). Indeed, Gordon and Harvey (2019) investigated the language of instruction favoured by different levels of South African society for the period 2003 to 2016, and found that a preference for English as the language of instruction grew from 55% to 65% during this period. This preference has been observed elsewhere (Civan & Coskun, 2016). Civan and Coskun (2016:1986) mention a particular example of a Tanzanian student who "refused instruction in Tanzania's local language Kiswahili, because he believed that English is the language of science and technology and modernization and development".

Most South African L2 students are likely to have been exposed to English from an early age, given the pervasiveness of English in the country. To what extent this is true of students who are from a deeply rural area, where access to technology and Internet connectivity are generally more limited than in the cities, is hard to assess without further investigation. For example, as reported in §1.2.1, Munje and Jita (2020) observed a lack of technology, for various reasons, in certain primary schools. Furthermore, despite the widespread presence of English, this is not the only additional language to which students are exposed, considering the variety of languages by which they are surrounded at any given time.

A second point on which South African L2 students differ from the L2 students described in international research is that, in South Africa, students do not typically share the same L1. Rather, within any classroom, the L1 is likely to be one of a selection of the eleven official languages. On the one hand, this has been shown to be an advantage in that it resulted in students resorting to English as the most efficient medium of communication (Scott, 2015:102). On the other hand, students have also been shown to be less expressive in their interpersonal communication in a group of mixed L1 students (Scott, 2015:102).

A final, though no less important, point to be mentioned with regard to the language issue in South Africa, and, in particular, the diversity of L1 languages, is that, "[i]t is to be expected that different varieties of English will develop in such an environment" (Wissing, 2002:129). Hence, the use of English, as discussed in this research and though drawing upon international research, will primarily be considered within the context of South African English.

In light of the complexities surrounding the learning of English by L2 students in South Africa, this research study aims to provide an indication of the success with which one particular feature of English is assimilated by L2 students through an investigation of PV use in academic writing.

A brief explanation of the language issues faced by South African L2 students was provided in this section. Next, an overview will be given of PVs, touching on how they are defined, and how they function in the English language.

1.3 Overview of PVs[1]

The English language is permeated by multiword structures (examples of which are idioms, such as *go the extra mile*, stock phrases, such as *thank you in advance*, and PVs, such as *call off*) (Gardner & Davies, 2007:339). Of these multiword structures, Gardner and Davies (2007:340) assess PVs as most significant for understanding English multiword structures, as

> [t]he study of phrasal verbs promises to provide valuable insights into what many linguists and applied linguists have begun to recognize as a multiword middle ground between "syntax and lexis" that has important ramifications for second language acquisition.

PVs constitute "a major grammatical class", according to Gardner and Davies (2007:347). As illustration, they state that PVs, "as a grammatical class, have a higher overall frequency than the verb

[1] Despite having similarities with English with respect to PVs, Afrikaans was not singled out for discussion from among the other African languages as an analysis of PVs within the various SA languages was not within the scope of the study.

are, the determiners *this* or *his*, the negative *not*, the conjunction *but*, or the pronoun *they*" (Gardner & Davies, 2007:347).

The use of PVs by L2 learners of English has been widely investigated internationally for a variety of mother-tongue speakers (Blais, 2012; Chu, 1996; Gaston, 2004; He, 2017; Kamarudin, 2014; Mazaherylaghab, 2015; McPartland-Fairman, 1989; Morales, 2000b; Qiu, 2018). The amount of research into PVs illustrates the extent to which they are regarded as key to the achievement of competence in English, and the degree to which they contribute towards L2 language users sounding "like native English speakers" (Haider *et al.*, 2020:1185). Furthermore, knowledge and use of multiword phrases such as PVs develop "learners' nativelike fluency" (Gardner & Davies, 2007:339) and "attest to mastery of English" (Riguel, 2014:112). Chen (2013a:420) maintains that the command that L1 writers have of PVs "is considered an important difference between their writing and learner writing". The correct use of PVs is, therefore, essential knowledge for the L2 student.

The PV consists of a *verb + particle* combination, with multiple verbs combining with a particular particle, and multiple particles combining with a particular verb, a unique PV forming on each occasion (McArthur, 1989:39). Each of these combinations might have multiple meanings, depending on the context in which they are used. Because of the unity inherent in the PV, several linguists who focus their research on vocabulary studies, such as Schmitt and Redwood (2011:174), view PVs as lexical items that are "holistic multi-word units" that are learnt as formulaic sequences, thus dispensing with the need for a more precise definition. Indeed, defining the PV has proved to be complicated, primarly because of the particle. Some researchers consider the particle as being either an adverbial particle or a preposition (Celce-Murcia & Larsen-Freeman, 1983; Darwin & Gray, 1999:69; Haugh & Takeuchi, 2023; Liao & Fukuya, 2004:196; Liu & Myers, 2020), whereas others view it as an adverbial particle only (Biber, Johansson, Leech, Conrad & Finegan, 1999; Chen, 2013a; Dagut & Laufer, 1985; Gardner & Davies, 2007; Hulstijn & Marchena, 1989; Trebits, 2009:471). The second version, that of a PV

consisting of a verb proper and an adverbial particle, is the one that has been adopted in this research study. The definition of the PV will be examined in detail in §2.3.

PVs have been found to be a difficult grammatical structure for second-language learners to acquire. PVs are semantically complex, as the meaning of the *verb proper + particle* combination cannot necessarily be derived from individual parts (Zarifi & Mukundan, 2014:52). Further difficulty results from the obscurity of idiomatic PVs, such as *bottle up*, *blow over* and *tune out* (Aldukhayel, 2014; Kamarudin, 2014; Mazaherylaghab, 2015). Additionally, a single PV might have various meanings, and meanings might also vary across registers (Chen, 2013b:89; Schmitt & Redwood, 2011:174). The PV may also be considered problematic because of its syntactic flexibility, in that the the two parts of the PV combination may be separated and still retain its unity (Gardner & Davies, 2007:341). The problems that the PV poses for the L2 learner may result in misinterpretation of the received message or avoidance of its use altogether (Darwin & Gray, 1999:65).

The most effective solution to this dilemma might be that of avoiding PVs altogether. Yet the consequence of such a course of action would be to acquire an English language that sounds awkward and unnatural. Rather, as Garnier and Schmitt (2016:30) maintain, the factors highlighted previously make it “all the more necessary to include them in the curriculum”, as they are “an important component of English vocabulary”. There is still lack of clarity as to the most effective way to teach PVs, although the emergence of corpus linguistics has facilitated their identification and classification (Gardner & Davies, 2007:342), and has provided researchers with a more complete view of student PV use (Chen, 2013b:91).

Previous research studies have approached PVs from different perspectives, including comparisons of the use of PVs in native and non-native writing (Mazaherylaghab, 2015; McPartland-Fairman, 1989; Qiu, 2018), the avoidance of PV use in L2 writing (Dagut & Laufer, 1985; Gaston, 2004; Haugh & Takeuchi, 2023; Laufer & Eliasson, 1993; Liao & Fukuya, 2004), the possible influence of a mother tongue without a similar language structure (Blais, 2012; Chu, 1996), and an investigation into PV use

by L2 speakers across both spoken and written registers, with particular reference to academic writing (Myers, 2018).

Various aspects concerning PVs were discussed in this brief overview. Next, the rationale for this study will be given.

1.4 Rationale for the study

As previously discussed, the correct use of PVs is a marker of having attained competence in English. Yet, to date, there seems to have been no in-depth research into the use of PVs in the South African context, and, furthermore, no comparative studies of L1 and L2 PV use appear to have been conducted. This fact will form the basis of the problem statement, as discussed in the next section.

The language difficulties pertinent to the South African educational context have been shown to be primarily concerned with L2 students, given the importance of English competence within an academic environment. While the PV use of L2 students will also form the main focus here, it cannot be viewed in isolation from L1 PV use. As will be seen in the literature review (particularly in §2.9), research into L2 PV use normally makes use of a native-speaker corpus (such as the Louvain Corpus of Native English Essays (LOCNESS)). However, in this research, L2 PV use will be compared to L1 student PV use from the same country (South Africa). There are several reasons that make this appropriate: first of all, it is not the aim in this research to investigate L2 PV use in isolation, but rather to obtain a comprehensive picture of student PV use in South Africa. This also means that a comparison of the two groups will not be focused only on differences, but also on the similarities that emerge from the investigation. Secondly, including South African L1 students in the investigation is suitable because of the unique nature of South African English, described by the *Oxford English Dictionary* (2023) as "firmly rooted in South Africa by the influence of the languages surrounding it". It is also interesting to note that "South Africans are often unaware of just how different [South African English] is from other Englishes in both vocabulary and pronunciation" (*Oxford English Dictionary*, 2023). L2 PV use should, therefore, be seen in the context of the L1 with which it is associated.

Having provided the rationale for the present research, the research aims and questions will now be expounded.

1.5 Research aims and questions

Although the primary focus of this research is on L2 student PV use, the purpose is to provide a picture of overall PV use. Therefore, both L2 and L1 PV use is investigated and reported on, using the same parameters of measurement throughout, which facilitates comparison between the two groups of students. As the academic texts of students at three South African institutions are available, it is also possible to compare PV use among these institutions in order to assess whether PVs are used differently at different universities. A possible preference for one-word alternative verbs over PVs is also investigated, as this further informs PV use. Non-standard use of PVs by both L2 and L1 students is reported on, as this adds an extra layer of understanding regarding the way the PV is employed in student academic writing.

Given the purpose of the research, the following research aims may be stated:

- To determine how L1 students use PVs in their writing.
- To determine how L2 students use PVs in their writing.
- To determine whether there is a difference in PV use between L1 and L2 students.
- To determine whether there is any difference in patterns of PV use across universities according to the ranking of the universities.
- To determine how patterns of PV use identified in the study could aid in the teaching of PV use to L2 students.

These research aims are addressed by means of related research questions which drive the research. The following group of questions focusses on the patterns of PV use in L1 student writing:

1a. Which phrasal verbs are predominantly used by L1 students?

1b. To what extent do L1 students show a preference for phrasal verbs or for one-word alternatives?

1c. To what extent do L1 students adhere to syntactic and semantic norms in their use of phrasal verbs?

A similar set of questions is then used to investigate the patterns of phrasal verb use in L2 student writing:

2a. Which phrasal verbs are predominantly used by L2 students?

2b. To what extent do L2 students show a preference for phrasal verbs or for one-word alternatives?

2c. To what extent do L2 students adhere to syntactic and semantic norms in their use of phrasal verbs?

Having investigated PV use from different viewpoints, the next question is used to address the overall aim of the research, followed by a question that is aimed at examining whether PV use at a university can be linked to its ranking.

3. What are the main differences in the use of phrasal verbs by L1 and L2 students?

4. Does the occurrence of phrasal verb use differ across universities according to the ranking of the universities?

The last two questions are aimed at addressing ways in which this study can aid in the teaching of PV use.

5. How could raising awareness of phrasal verbs and their alternative one-word verbs help students make appropriate choices in their academic writing?

6. How could reviewing errors in the use of phrasal verbs serve to guide students on how to use phrasal verbs correctly?

The research aims and questions that will drive this research study were expounded in this section. Hereafter, an overview will be provided of the methodology to be used, including the research approach and theoretical framework. The corpora to be used are briefly described, as well as the participants, and the parameters within which PV use will be investigated.

1.6 Overview of methodology

This research uses a corpus-based approach in which a research question is identified as a starting point (Rayson, 2008:519), and patterns of language use in large corpora are investigated by means of suitable software (Biber, Conrad & Reppen, 1998:4). In this case, the software selected was Wordsmith Tools (Scott, 2022) as it enables the generation of frequency lists and concordance lines. As WordSmith Tools delivers PV frequency tables, quantitative techniques are required for the analysis of the results. At the same time, qualitative techniques are necessary to interpret the patterns of PV use that are observed. Corpus linguistics serves to provide the methodological approach because of the necessity of analysing language using a scientific method (Brezina, 2018:2), and because it allows for "greater generalizability and validity than would otherwise be feasible" (Biber *et al.*, 1998:1). Using descriptive research as the research design allows for an accurate description of the observed phenomena (Atmowardoyo, 2018:198).

Two corpora are used, one comprising student data from one institution only, namely the University of the Witwatersrand, and the other incorporating the student data of two institutions, namely North-West University and the University of Pretoria. The data from each institution will be separated into L1 and L2 data, forming L1 and L2 subcorpora within each institution, in order to adequately address the research questions. The participants are students at various stages of their degrees, and the student writing represented in the corpora consists of academic texts from a range of disciplines, covering various topics. While the L2 subcorpora represent a variety of South African languages, it

does not fall within the scope of this research to report on these subgroups (see §5.6), as the main aim here is the consideration of student writing within an English-medium academic context.

The extracted PV data are viewed from various angles, as PV frequency alone will not provide a fully rounded perception of how this grammatical construction has been applied in student writing. Therefore, not only will the number of times PVs occur per subcorpus be examined, but also the number of times a particular PV occurs in the subcorpus. An investigation into the distribution of PVs within a subcorpus will provide further insight into how a particular group of students use PVs. To expand on the information garnered from the patterns of PV distribution, the ten most frequently used PVs per subcorpus will be extracted, and the presence of one-word alternative verbs to these PVs in the subcorpora will then be investigated to ascertain whether students reveal a preference for either form. During the process of PV validation, non-standard use of PVs will be noted and reported on to provide additional information about PV use.

This section gave a short overview of the methodology to be used in the research. Finally, in the next section, the structure of the thesis will be presented as a guide to the reader.

1.7 Structure of the thesis

The thesis is presented in a traditional five-chapter structure. The introductory chapter will be followed by the literature review chapter. This chapter provides background to the research, firstly, by explaining the history of how the PV came into being, and how this resulted in some of the characteristic difficulties for which it is known. Secondly, the definition that will be used for the PV throughout this research is explained, which necessitates enumerating the complications involved in defining the PV. Further information is provided regarding the nature of PVs, such as the various categories into which PVs fall, and how PVs are used across registers. Some of the many studies that have investigated frequently used PVs are then discussed. Thereafter, an explanation is given of some of the issues that L2 students face in the use of PVs, and, finally, various research studies into

avoidance, which is the primary means by which students attempt to overcome these issues, are described.

The methodology chapter expounds the methods to be used for the study. The problem statement that drives the research is clarified, and the research questions explained. The theoretical framework and research design most appropriate to the successful fulfilment of the proposed research are explained, as is the rationale, which clarifies the reason for the research being undertaken. As a pilot study was carried out before the start of the main research to ensure that possible problems with the research design could be resolved beforehand, the procedure followed for that study is explained, and the results presented. Next, the corpora to be used in the main study are described, as well as the participants. Both the quantitative and qualitative aspects of the data collection process follow, and, finally, the process used to analyse the results completes the chapter.

Chapter 4 covers results and analysis. The processes discussed in the methodology chapter are put into practice here, firstly, by extracting data on L1 and L2 PV use per institution, and reporting on the PV use according to the parameters given in §1.3. PV use across the three institutions is then investigated, followed by an overall comparison of similarities and differences between L1 and L2 PV use. The results are then discussed by observing the patterns of PV use that have emerged, and what these patterns of PV use signify more broadly regarding the way in which L1 and L2 students use English in their academic writing.

In the concluding chapter, the aims and research questions of this research study are revised, which leads into a discussion of the main findings. This discussion looks at the broader implications of the results that emerged from the investigation described in the previous chapter. It is then suitable to suggest what the contribution is that the research has provided, and, further, to propose pedagogical implications and make recommendations in this regard. In keeping with a research study of this nature, the limitations of the research are explained. The chapter concludes with suggestions for future research.

1.8 Conclusion

This chapter introduced the aim of this research, which is to investigate and report on PV use by South African L1 and L2 students. A background to the study was provided to substantiate the intended contribution of this research to the field of academic writing. Furthermore, an overview was provided of PVs as an initial clarification of their characteristics, and the issues involved in their use, leading to an explanation of the rationale for the study, the research aims and questions, and the methodology that is to be followed.

In the next chapter, the PV will be discussed in detail in order to emphasise the importance of its use in English, as well as the difficulty its use entails for L2 students. In addition, various relevant research studies into PV use globally will be discussed to provide background to the proposed research.

Chapter 2: Literature review

2.1 Introduction

At the start of any research, it is necessary to provide the framework in terms of which the research is to be conducted. As this study will focus on the use of PVs, it is important to start by giving a definition of the PV, as well as by providing further information about PVs, such as PV types, most frequently used PVs, and PV use across registers.

Furthermore, to validate the supposition that one's research will add value to an existing field of knowledge, one needs to acknowledge and draw on the theories, problems and findings of previous research studies. These studies will be grouped according to their focus, which will also provide a useful overview of the main areas of discussion surrounding PVs.

2.2 Brief overview of the origins of the PV

The origins of the PV can be found in the Old English period (± 450 AD to ± 1100 AD) during which time this verb form seems to have arisen naturally from everyday speech. While PVs of the *verb + adverb* form (for example, *ahof up*) already existed at this stage, Bolinger (1971:xi) detects only "a trace" of these combinations, and, likewise, Kennedy (1920:11) finds occurrences of these forms to be "practically nil". In fact, the more common construction for the PV was an *adverb + verb* combination, the adverb thus taking on a prefix position (Bolinger, 1971; Kennedy, 1920), such as *upstigon* (*climb up*), *upahafen* (*lift up*), and *utscufon* (*shove off*) (Kennedy, 1920:12). (There seems to be a difference of opinion among researchers as to whether the adverb is separable from the verb (Bolinger, 1971:xi), or inseparable (Kennedy, 1920).)

The *verb + adverb* form of the PV (for example, *smit of*, *wende ut*, and *etbrec ut*) gradually started to gain popularity in the Middle English period (± 1100 AD to ± 1500 AD) (Bolinger, 1971; Kennedy, 1920; McPartland-Fairman, 1989), although, at the start of this period, Kennedy (1920:12) still finds ample evidence of the predominance of the older (*adverb + verb*) form. However, further development of

both forms of the PV was slowed – at least in written work – because of the incorporation of Latinate (Romanic) word forms into the English language (Bolinger, 1971; Kennedy, 1920; McPartland-Fairman, 1989). The invasive Latinate words had an inseparable *prefix + verb* construction (such as *exstinguo (extinguish)* and *circumvenio (circumvent))*, and were somewhat similar in appearance to the original form of the PV. As with other Latinate words gaining ascendancy over Anglo-Saxon words during this time, these compound Latinate words were considered superior to PVs and were, therefore, preferred in formal writing, so that they "drove out the native [*adverb + verb*] compounds, and for a time made the newer [*verb + adverb*] combinations unnecessary" (Kennedy, 1920).

By the end of this period, and moving into the Modern English period (± 1500 AD to the present), the PV, in the form by which it is now known, started to reappear (Bolinger, 1971; McPartland-Fairman, 1989). In *The Ballad of Robin Hood and The Monk* (± 1450), Kennedy (1920:13) finds 20 occurrences of the newer *verb + adverb* form, 18 of the old *adverb + verb* form (see examples above), and none of the Latinate words, although he cautions that the work was intended to illustrate everyday English, as typically used in a social context, and that more formal works do not show the same pattern of PV use. This suggests that the PV had continued to be used in everyday speech by the "working man", and that its "reappearance" refers only to its presence in written work. As Kennedy (1920:40) astutely observes, "[p]erhaps the common man is not to be blamed for avoiding the use of a vocabulary which has never really been his, and for utilizing in the expression of certain ideas his own familiar stock of words". Because of this history, McPartland-Fairman (1989:1) refers to PVs as "Anglo-Saxon combinations", and to the one-word alternative as "single-word Romance equivalents". As for the *adverb + verb* construction, it eventually lost ground in favour of the *verb + adverb* PV form, although a few examples (such as *foreshadow*, *outnumber* and *overthrow*) are still in use today.

As a result of extensive new research by Elenbaas (2007) and Thim (2012), among others, a slightly more complicated process of the development of the PV has been suggested. According to Elenbaas (2007:3), a system of separable and inseparable complex verbs co-existed in Old English (± 450 AD to

± 1100 AD), both with a *prefix + verb* construction. In both cases, the prefixes functioned as "preverbs", forming a close association with the verbs that they accompanied (Elenbaas, 2007:105). Old English (± 450 AD to ± 1100 AD) was not unique in having preverbs: many languages of that time (± 500 AD to ± 1100 AD) presented parallel constructions. For example, Thim (2012) points out that the *prefix + verb* construction was typical of all Germanic languages. Furthermore, the evolution of preverbs shows similar mechanisms at work in their development across languages, although not necessarily at the same time (Elenbaas, 2007:105, 107).

While there were similarities between the separable and inseparable complex verbs, their differences were the primary drivers of their further development. In the case of the inseparable complex verb, the prefix, having a vague meaning, was dependent on and, accordingly, inseparable from the verb (Elenbaas, 2007:105, 133). As Thim (2012:104) puts it, the prefix was "fused to the verbal stem". In the case of the separable complex verb, the prefix, having an evident meaning, could be separated from the verb by grammatical insertions such as a negative marker or a modal verb (Elenbaas, 2007:105). In present day terms, this type of prefix would be considered a particle (by which term it will be identified in the rest of this discussion). The preverb position of this particle was the result of word order in Old English (± 450 AD to ± 1100 AD), which was far more flexible than the English of today. Consequently, as the relationships between words were marked by inflection and not by position, a variety of object-verb word orders were possible, as well as some verb-object word orders (Elenbaas, 2007:107-108).

In many instances, the particles and prefixes had comparable meanings, as can be seen in examples in current use, such as *light* ***up*** and ***il****luminate* (McPartland-Fairman, 1989:34), *blow* ***out*** and ***ex****tinguish*, and *get* ***around*** and ***circum****vent* (Bolinger, 1971:xi) (in these examples, the particle and prefix, respectively, are in bold). The vague nature of prefix meanings meant that they eventually became "too abstract to convey the intended meaning", so that particles began to be added to the inseparable complex verb to compensate for the lost meaning (Elenbaas, 2007:114-115). (As

illustration, Bolinger (1971:xii) provides present-day examples such as *extend out*, *refer back* and *proceed forth*.) Thus, during the Middle English period (± 1100 AD to ± 1500 AD), the separable complex verb started to dominate, and eventually eclipsed the inseparable complex verb (Elenbaas, 2007:173; Thim, 2012:111).

Early on in the Middle English period (± 1100 AD to ± 1500 AD), a further change occurred in that the particle moved from its preverb position to a postverbal position, possibly as a result of the change from an object-verb word order to a verb-object word order (Elenbaas, 2007:211; Thim, 2012:103). Such changes to language structures, such as word order, followed similar routes in both English and other Germanic languages (Thim, 2012:115). Thim (2012:116), therefore, calls PVs "typically Germanic". The particle became more dependent on the verb, and verb and particle combinations began to display the "unit-like behaviour" of the PVs we have today (Elenbaas, 2007:280). Particle meanings also expanded during this time, with the inclusion of less transparent, more idiomatic meanings (Elenbaas, 2007:216). From this point onward, the PV started to function in its current form.

2.3 Defining the phrasal verb

In the English language, lexical verbs (full verbs) comprise either single word or multi-word verbs that function as single lexical units (Biber *et al.*, 1999:358). Biber *et al.* (1999:403) identify three main categories of multi-word verbs, namely phrasal verbs (PVs) (e.g. *put on*), prepositional verbs (PrepVs) (e.g. *think about*), and phrasal-prepositional verbs (e.g. *put up with*). Celce-Murcia and Larsen-Freeman (1983:267) see the phrasal-prepositional verb as a sub-category of phrasal verbs. To the three categories, Biber *et al.* (1999:403) also add a fourth, into which they group three "major types of idiomatic multi-word verb constructions": *verb + prepositional phrase* combinations (PrepPs) (e.g. *keep in mind*), *verb + verb* combinations (e.g. *make do*), and *verb + noun phrase* combinations (e.g. *make peace with*) (Biber *et al.*, 1999:427-428). Of all the combinations identified by Biber *et al.* (1999:403), this study is concerned primarily with the phrasal verb (PV), although the prepositional verb (PrepV), and *verb + prepositional phrase* (PrepP) will also be discussed because of the necessity

of differentiating between the PV and these two combinations. The phrasal-prepositional verb, on the other hand, is easily distinguishable from the PV because of its structure (*verb + particle + preposition*), as is the *verb + verb* combination, as well as the *verb + noun phrase* combination.

Before discussing the definition of a PV, there are some general characteristics of the PV that should be mentioned. PVs are seen as semantic units as they function as single verbs to all intents and purposes. Gilquin (2015:55) describes the verb and particle as forming "a single unit of sense in which the particle modifies or completes the meaning of the verb". This inherent unity is illustrated by the following characteristics: they can be transitive, intransitive, or ergative; they can form the passive voice and action nominals; and they are stressed on the last syllable (Darwin & Gray, 1999:69). The only aspect in which they differ from single verbs is in the ability for their two parts to be separated by the insertion of the direct object of the sentence between the simple lexical verb (verb proper) and the particle (Darwin & Gray, 1999:69). This aspect is significant and will be discussed in more detail in §2.3.1.

Furthermore, the simple lexical verb that forms part of the PV is mostly monosyllabic (e.g. *take*, *get*, *put*, *come*, and *go*), although a limited number of PVs have bisyllabic simple verbs, such as *divide up*, and, very rarely, trisyllabic simple verbs, such as *partition off* (McPartland-Fairman, 1989:33-34). McPartland-Fairman (1989:34) notes the Teutonic origin of the monosyllabic simple verbs, which points to their Anglo-Saxon derivation.

A simple lexical verb can be joined to a range of adverbial particles, each combination forming a unique PV (e.g. the lexical verb *turn* can become the PVs *turn up*, *turn out*, *turn on*, *turn in*, *turn off*, *turn over*, and *turn down*) (McArthur, 1989:39). A further important aspect of PVs is that a single PV can have several meanings (e.g. *make up* can refer to *constituting a group*, *restoring a relationship*, *creating a story*, *mixing medicinal ingredients together, preparing a bed*, or *completing a required number*) (Biber *et al.*, 1999:408; Bronshteyn & Gustafson, 2015:92). These aspects indicate the ease with which new combinations are formed and highlight the highly prolific nature of PVs, with new PVs regularly being

added to the English language (e.g. *man up*, *psych out*, *rev up*, *log in,* and *zone out*) (Bronshteyn & Gustafson, 2015:92; Riguel, 2014:112). (Bolinger (1971:xiii) makes the interesting observation that PVs also generate a myriad of new nouns, and gives *standoff*, *runaway* and *makeup* as examples.) Bolinger (1971:xii) ascribes the prolificacy of the PV to "the familiarity and manageability of the elements". PVs are, therefore, a common feature of English generally (Bronshteyn & Gustafson, 2015:92; McPartland-Fairman, 1989:1).

Darwin and Gray (1999:68) rightly claim that "to avoid an ambiguous classification procedure, linguists must agree upon a definition, thereby requiring them to begin from the same point". However, there does not seem to be a universally accepted definition of the PV among researchers (Chen, 2013a; Dagut & Laufer, 1985; Darwin & Gray, 1999; Gardner & Davies, 2007; Hulstijn & Marchena, 1989; Liao & Fukuya, 2004; Liu & Myers, 2020; Wilcoxon, 2014). In fact, Darwin and Gray (1999:66) suggest that the absence of a transparent definition is one of the reasons for "the lack of progress in the understanding of phrasal-verb pedagogy". McArthur (1989:39) suggests another interesting reason, that as PVs "have for centuries been part of that 'plain' foundation underneath the French and Latin superstructures of the language, they have attracted little attention among classically-inspired grammarians", and have, therefore, been "confined to the fringes of grammatical description".

The detailed dissecting of the PV, often necessarily accompanied by other multi-word combinations, has led to a confusion of terminology regarding this grammatical structure. Compare, for example, the contrasting views on PVs in publications such as the *Longman Grammar of Spoken and Written English*, where the PV is given as *a verb + an adverbial particle* (Biber *et al.*, 1999:407-413), as opposed to *The Grammar Book: An ESL/EFL Teachers' Course*, where, although the PV is also given as *a verb + a particle*, the particle is "variously described as a preposition, an adverb or a combination of the two" (Celce-Murcia & Larsen-Freeman, 1983:265). Furthermore, Bolinger (1971:6) doubts that "a linguistic entity such as the phrasal verb can be confined within clear bounds", and adds that "being or not being a phrasal verb is a matter of degree". Under these circumstances, there is justification in the oft-

quoted comment by Gardner and Davies (2007:341) in which they ask: "if even the linguists and grammarians struggle with nuances of phrasal verb definitions, of what instructional value could such distinctions be for the average second language learner?"

The search for a valid yet practical definition of the PV has, therefore, become the necessary pursuit of researchers in this field, of which the two discussed here are the most current examples (Darwin & Gray, 1999; Gardner & Davies, 2007). Darwin and Gray (1999:68) define the PV as "a verb proper and a morphologically invariable particle that function together as a single unit both lexically and syntactically". On the other hand, Gardner and Davies (2007:341) describe PVs as "all two-part verbs […] consisting of a lexical verb proper […] followed by an adverbial particle […] that is either contiguous (adjacent) to that verb or noncontiguous (i.e., separated by one or more intervening words)".

In both cases there is agreement that the PV (being a *two-part verb)* functions as a unit, whether it is functionally separated by the insertion of another word or not. While the "concept of separation" is not mentioned in their definition, Darwin and Gray (1999:77) do emphasise its use in clarifying their definition. Both definitions allow for "varying degrees of semantic transparency" (Gardner & Davies, 2007:341), in contrast to the insistence of, for example, Biber *et al.* (1999:403) that PVs "cannot be derived from the individual meanings of the two parts", and the even more emphatic insistence by some researchers, such as Bolinger (1971) and McPartland-Fairman (1989), that a PV that can be taken literally cannot be considered a PV at all.

The description that Darwin and Gray (1999:69) provide of the particle as "morphologically invariable" is useful in distinguishing PVs from multiword combinations where the verb combines with personal pronouns, adjectives or nouns. This limits the particle to being a spatial adverb (or adverb of place) or a preposition (Darwin & Gray, 1999:70), although this limitation by no means concludes the argument. The main point of contention regarding the identification of PVs, therefore, revolves around the particle that follows the verb proper. Consequently, we see that some researchers advocate for only an adverbial particle after the verb proper (Biber *et al.*, 1999; Chen, 2013a; Dagut & Laufer, 1985;

Gardner & Davies, 2007; Hulstijn & Marchena, 1989; Trebits, 2009:471), while others include a preposition in the definition (Celce-Murcia & Larsen-Freeman, 1983; Darwin & Gray, 1999:69; Haugh & Takeuchi, 2023; Liao & Fukuya, 2004:196; Liu & Myers, 2020).

The issue is by no means straightforward, as an attempt to distinguish between adverbial particle and preposition will illustrate. For example, most particles mentioned in PV research can function as either adverbs or prepositions (Biber *et al.*, 1999; Celce-Murcia & Larsen-Freeman, 1983; Chen, 2013a:424; Darwin & Gray, 1999; Liu & Myers, 2020), and can only be identified as either by their role in a sentence. The following definition given in *The Cambridge Grammar of the English Language* (Huddleston, Pullum & Bauer, 2002:603) supports the idea that the particle should be viewed more broadly:

> a relatively closed grammatically distinct class of words whose most central spatial members characteristically express spatial relations or serve to mark various syntactic functions and semantic roles.

The next two sections will be focused on defining the PV, and will conclude with the PV definition to be used in this study.

2.3.1 Defining the PV as *verb + adverbial particle or preposition*

Let us consider the view of those who include both spatial adverb and preposition in the definition of a PV. While Celce-Murcia and Larsen-Freeman (1983:265) use this framework they do not, in fact, refer to these identifiers at all, seeing the particles rather "as being similar to but not identical to prepositions", and calling them simply *particles*, "i.e., a new part of speech distinct from adverbs or prepositions". Celce-Murcia and Larsen-Freeman (1983:266) mention two specific characteristics of PVs: firstly, that PVs may be either transitive ('*He had to **fill out** a form*'[SI 2]) or intransitive ('*You should **come over***'[SI]) ; and, secondly, some of the transitive PVs may be separated ('*She **turned** the stove **on***'[SI])

[2] All example sentences with the superscript[SI] are my own. Other example sentences are taken from the referenced source.

or not (‘*She **turned on** the stove’*), while others are inseparable (‘You should *go over* your notes’[SI] – not: *’*You should **go** your notes **over**‘) or require separation (‘*I need to **get** the idea **through** to him*’[SI], in contrast to the inseparable PV in *’I don’t know how I **got through** the day*’[SI]. Of these two specific characteristics, the first applies to either definition of the PV, and the second applies only to the definition that includes both the spatial adverb and preposition. This is because Celce-Murcia and Larsen-Freeman (1983:268-270) include inseparable PVs (‘*You should **go over** your notes*’[SI]) and PVs that must be separated (‘*She should **see** the job **through** to completion*’[SI]) in their definition of the PV, which differs from the alternative PV definition, as discussed in the next section. They acknowledge that “some grammarians do not distinguish between verb-plus-preposition sequences and inseparable phrasal verbs”, but consider this lack of distinction to be a mistake (Celce-Murcia and Larsen-Freeman (1983:268-270). When using the *verb + adverbial particle or preposition* definition, the PV needs to be distinguished from a PrepP, but not from the combinations that Celce-Murcia and Larsen-Freeman (1983:268-270) identify as inseparable PVs (referred to in this research as PrepVs). The aim, therefore, is not necessarily to differentiate between particles and prepositions, as the PrepV essentially has a preposition following the verb proper.

Celce-Murcia and Larsen-Freeman (1983:268-270) suggest three tests aimed at distinguishing between PVs and PrepPs , all of which relate to the unity of the PV. The first test, called “unacceptable fronting in wh-questions”, makes use of wh-questions to see if the particle can be moved before the verb (‘*She **turned off** the stove*’[SI] cannot be changed into *’***Off** *what did she **turn***?’, therefore, *turn off* is a PV; in contrast, *’She **looked at** the view*’[SI] can be changed into *’**At** what did she **look**?*’, and, therefore, *look at* is a PrepP). The second test, known as “unacceptable fronting in relative clauses”, looks at whether the sentence can be turned into a relative clause (**’The stove **on** which she **turned**…’ is not acceptable, which means that *turn on* is a PV; in contrast, *’The view **at** which she **looked**…*’ is acceptable, which shows that *look at* is a PrepP). The third test inserts an adverb between verb and particle (**’She **turned** quickly **on** the stove*‘[SI] is not acceptable, which confirms that *turn on* is a PV; in contrast, *’She **looked** approvingly at the view*‘[SI] is acceptable, showing that *look at* is a PrepP). The

validity of this third test has, however, been queried because of the existence of PV-containing sentences such as: 'He *walked* quickly *away'* (McPartland-Fairman, 1989:39).

In the next section, the definition of a PV as a *verb + adverbial particle* will be discussed.

2.3.2 Defining the PV as *verb + adverbial particle* only

Let us now consider the view of those who advocate for limiting the particle to an adverbial particle. As indicated previously, the adverbial particle that is used in a PV is identified as a spatial adverb (Darwin & Gray, 1999:70), having "core spatial and locative meanings" (Biber *et al.*, 1999:403), although Biber *et al.* (1999:403) concede that it is used "with extended meanings" in this case.

Limiting the PV to a *verb + adverbial particle* requires a change in the way a PV is differentiated from other multi-word verb combinations. In the previous section, where the definition of the PV was *verb + adverbial particle or preposition*, the PrepV, constituting a *verb + preposition*, was given as an inseparable PV. It, therefore, did not need to be differentiated from a PV. However, as the PV definition now excludes the preposition, differentiation will be needed between PVs and PrepVs, along with other multiword combinations.

It is necessary to have some understanding of the PrepV, inasmuch as it needs to be distinguished from the PrepP. A characteristic of a PrepV is that it is always transitive, and, consequently, always has a direct object (Wilcoxon, 2014:8). This characteristic could make it challenging to distinguish a PrepV from a PrepP when a PrepP consists of a noun or noun phrase. Compare, for example, '*I need to go* ***for a haircut***'[SI] (PrepP) with '*The dog is likely to* ***go for*** *the* cat'[SI] (PrepV). McArthur (1989:42) does not recognise a grammatical structure such as the PrepV, calling all *verb + preposition* constructions simply "straightforward prepositional phrases". Nevertheless, he does recognise that situations exist where the *verb + preposition* have a closer link than usual, which he illustrates with "*She* ***came across*** *an old friend*" in contrast to "*He* ***came across*** *the street*" (these two examples are taken from McArthur (1989:42)). He describes the situation as "stealing" the preposition from the PrepP and "fusing" it with

its own verb to form a new idiomatic relationship. Consequently, his own term for this type of combination is "a *fused or non-separable phrasal verb"* [author's emphasis] (McArthur, 1989:42).

In order to clearly distinguish between the PrepV and PrepP, a second characteristic of the PrepV has special relevance: the object of a PrepV will answer "who" or "why" questions, whereas the object of a PrepP will answer "where" and "when" questions (Biber *et al.*, 1999:405; Wilcoxon, 2014:8). This becomes a useful test to distinguish a PrepV from a PrepP when a PrepP consists of a noun or noun phrase (Biber *et al.*, 1999:406). For example, for the statement '*The most important point **came from** the Nigerian speaker*'[SI], the question '*Who did the most important point **come from**?*' can legitimately be asked (therefore, a "who" question, servicing a PrepV), which means that the phrase *came from* is a PrepV and *the Nigerian speaker* is the direct object. On the other hand, for the statement '*The keynote speaker **came from*** Nigeria'[SI], the question '*Where did the keynote speaker **come from**?*' can be asked, which makes the phrase ***came from** Nigeria* a PrepP.

It should be noted that Biber *et al.* (1999:403), moreover, add "free combinations" to the mix, "where each element has a separate grammatical and semantic status", meaning that the verb is followed by a particle, or by an adverbial prepositional phrase, that carries its own meaning. This definition is clearly intended to differentiate the free combination from PVs (and PrepVs) since they, in contrast to free combinations, "represent semantic units that cannot be derived from the individual meanings of the two parts", as Biber *et al.* (1999:403) insist. The distinction, however, is tenuous. For one thing, the requirement of PV idiomaticity is no longer strictly held, as will be seen in the discussion on the classification of PVs. In fact, Gardner and Davies (2007:341) do not take "degrees of semantic transparency" into consideration at all when identifying PVs. A further point is the difficulty of assessing to what extent "individual meanings" can be assigned to the *verb + particle*. Indeed, Biber *et al.* (1999:403) admit that,

> In practise [*sic*], it is hard to make an absolute distinction between free combinations and fixed multi-word verbs; one should rather think of a cline on which some verbs, or uses of verbs, are relatively free and others relatively fixed.

The example provided by Biber *et al.* (1999:403) of a free combination consisting of a *verb + particle*, namely *go back*, is a case in point. In the *Merriam-Webster Online Dictionary* (2024) and *Cambridge Phrasal Verbs Dictionary* (2006) it is identified as a PV, yet is given as an intransitive free combination by Biber *et al.* (1999:403), and contrasted to intransitive PVs (e.g. *fall in*, *fit in*, *come in*, *stay on*). However, Biber *et al.* (1999:403) admit that the categorisation of multi-verb combinations often depends on context, so that *stay on* in '*I would like to **stay on** and honour my contract*' is an intransitive PV, while *stay on* in '*Many dealers were content to **stay on** the sidelines*' is a free combination of *verb + prepositional phrase* (These two examples are from Biber *et al.* (1999:406)). The free combination category does not appear to be widely used in other research, and will not be held to here.

As in the previous section, the definition of the PV currently under discussion requires differentiating the PV from all other multi-word verb combinations, but with the added qualification that the multi-word verb combinations now include the redefined PrepV. All prepositional combinations, consequently, can now be tested for and rejected (Biber *et al.*, 1999:403; McArthur, 1989:42; Wilcoxon, 2014:7-8). The most effective test for distinguishing PVs from PrepVs (and, effectively, from all other multi-word verb combinations) is particle movement (Biber *et al.*, 1999:404; Wilcoxon, 2014). This is because what is most often true of a transitive PV is that the direct object may be positioned between the verb and particle (mid position), as in '*She **picks** the book **up**'* [SI], or after the particle (end position), as in '*She **picks up** the book*' [SI]. Such separation is not possible for the PrepV. For example, using a previous illustration, the verb and particle of the PrepV in the sentence '*The most important point **came from** the Nigerian speaker*'[SI] cannot meaningfully be separated: *'*The most important point **came** the Nigerian speaker **from***'[SI]. This points to "the clearest syntactic difference between particles and prepositions: only particles can occur after the direct object" (Morales, 2000a:16). According to Biber *et al.* (1999), the more idiomatic the PV, or the more formal the writing, the less comfortable verb and particle separation may seem. On the other hand, in the case of the direct object being a pronoun, the choice of position disappears: the pronoun can then only be placed between the verb and particle, as in '***Pick** it **up***' (Biber *et al.*, 1999:408; McArthur, 1989:39). In fact, pronoun

placement constitutes one of the tests used to differentiate between PVs and PrepVs, as will be seen below.

It should be pointed out that the particle placement test applies only to transitive PVs. However, intransitive PVs (e.g. *sit down*, *grow up*, *find out*) can be taken at face value, as, not having a direct object, they cannot be confused with any other multi-word verbs, which all require direct objects. As discussed previously, Biber *et al.* (1999:404) do make a distinction between an intransitive PV and a free combination (*verb + adverb*, where both retain their independent meaning), but this difference was shown to be vague and difficult to apply.

Morales (2000a) adds several other tests that are intended to facilitate the distinction between particle and preposition. Of these tests, there are three with which only prepositions would be able to comply. For example, he suggests using "phrase fronting" to identify PrepVs, as PVs should fail the test. Therefore, *'***On** *the stove she* ***turned**'*, which is not acceptable, can be identified as a PV. Nevertheless, this is not a completely reliable test for identifying PrepVs because of the often unnatural result (?'***At** *the view she* ***looked***'). On the other hand, the "inseparable PV" that Celce-Murcia and Larsen-Freeman (1983:268-270) include in the PV category, but which is identified as a PrepV here, also passes this test because of its *verb + preposition* structure ('***Across** the bridge they* ***came***' is acceptable, which means that *come across* is a PrepV). Two of the tests used by Celce-Murcia and Larsen-Freeman (1983:268-270) are also mentioned by Morales (2000a:16), namely unacceptable fronting in wh-questions, and insertion of adverbs (illustrated in the previous section). As was seen in the discussion regarding these tests, they do not detect all prepositions, and, therefore, further tests are required to differentiate between particles and prepositions.

The other tests suggested by Morales (2000a:16) should only be successful for PVs. The first of these tests is the forming of the passive voice which should be possible for PVs ('*The stove was* ***turned on***' is acceptable, making *turn on* a PV) but not for PrepVs (*'*The bridge was* ***come across***' is unacceptable, as it is a PrepV). However, '*The view was* ***looked at***' is likewise acceptable, even though *looked at* is a

PrepV, which suggests that this test is, again, of dubious value. The verb substitution test claims that, in contrast to a PrepV, it is possible for a PV to be replaced by a single word, since "verb substitution is an exclusive property of particles" (Morales, 2000a:16), which means that the PrepV would automatically be excluded. Yet, here, too, the test proves inconclusive, as not all PVs have synonyms (Chen, 2013a:436; Darwin & Gray, 1999:71). For example, there is no one-word substitution for the PV in '*She **switched on** the stove*'. On the other hand, there is a one-word substitution for the PrepV in '*She **looked at** the view*', namely *observed*, as well as for '*They **came across** the bridge*', namely *traversed*. (One-word substitutions for PVs will be discussed in more detail in §3.8.2.5.)

Many of these tests (particle placement, adverbial insertion, formation of passives, verb substitution) correspond to the nine PV verification tests reported by Bolinger (1971), although he concurs with the weaknesses that have been pointed out. He also lists the formation of action nominals, pronoun placement, PV stress, and PV listing as useful tests for PV verification, each of which will briefly be explained. The formation of action nominals refers to deriving a noun from a transitive PV (Bolinger, 1971:18; Darwin & Gray, 1999:71). Thus, this construction should be possible for a PV ('*Her **turning on** of the stove*' is acceptable, meaning that *turn on* is a PV), but not for a PrepV (*'*Her **looking at** of the view*' is not acceptable, making *look at* a PrepV). Likewise, the inseparable PVs included in the PV category by Celce-Murcia and Larsen-Freeman (1983:268-270) fail this test (*'*Her **coming across** of a bargain*' is unacceptable, making *come across* a PrepV). The pronoun placement test requires the direct pronoun to be placed between the verb proper and the particle (Bolinger, 1971:11; Darwin & Gray, 1999:73), which should be possible only for the PV ('*She **turned** it **on***' is acceptable, because *turn on* is a PV; *'*She **looked** it **at***' is not acceptable, because *look at* is a PrepV). Bolinger (1971:11) considers this one of the easiest tests to apply. The third test in this list specifies that a PV can be identified by the use of stress. For example, stress will naturally be placed on both parts of the PV in the sentence '*She **túrned ón** the stóve*', whereas, in the case of a PrepV, the stress will be placed on the verb proper only ('*She **lóoked at** the víew*'). Although this differs somewhat from the contention by Darwin and Gray (1999:69) that stress is mostly placed on the particle, there is concurrence in the

fact that, in both cases, the particle remains unstressed in the case of a PrepV. According to Darwin and Gray (1999:73), this is because "prepositions, not being content words, do not receive stress". Finally, a straightforward method of identifying PVs, and "not a test at all" (Darwin & Gray, 1999:74), is to simply list them. Bolinger (1971) identifies two problems with this suggestion: firstly, the impossibility of creating an exhaustive list for such a productive grammatical category; and, secondly, the difficulty of adequately incorporating regional differences. Nevertheless, such attempts have been made, and to good effect (Gardner & Davies, 2007; Garnier & Schmitt, 2015; Liu, 2011; Liu & Myers, 2020; Myers, 2018; Schmitt & Redwood, 2011), as will be seen in the discussion on frequency of PV use (§4.6).

Darwin and Gray (1999), in pointing out the exceptions that slip through each test, illustrate the insufficiency of the tests identified above. In an attempt to simplify PV identification, Darwin and Gray (1999:76) suggest an alternative method that allows researchers "to throw out rather than to throw in": all *verb + particle* combinations are taken to be PVs, unless the combinations fail any one of seven tests that they have devised (Darwin & Gray, 1999:77-81). This method, however, has also been criticised for its shortcomings, as they exclude PV forms accepted elsewhere. For example, their test of the acceptable insertion of two adverbs ending in "-ly" being proof that a combination is not a PV would exclude the PV *walk out* (Sawyer, 2000), because of the acceptability of '*They* ***walked*** *quietly and quickly* ***out***'. Furthermore, it is felt that their classification system does not significantly add to the field (Gardner & Davies, 2007:341).

This discussion on the suggested means of separating PVs from other multi-word verbs (especially the PrepV) suggests that, as Biber *et al.* (1999:405) and Darwin and Gray (1999:71-75) point out, distinguishing among the multi-word verb categories is not always clear cut. In fact, "several verb combinations can function as more than one type, depending on the context; and some particular combinations can be interpreted as belonging to more than one category". According to Chen (2013b:90), researchers largely acknowledge that "whatever tests are used, there are always

noticeable exceptions". She also points out the time wastage of having to check *verb + particle* combinations against a multitude of tests (Chen, 2013b:90). Furthermore, the variety of viewpoints as to what constitutes a PV or a particle can lead to confusion, and are "of little instructional value for non-native speakers and learners of English struggling to master this area of vocabulary and grammar" (Trebits, 2009:471).

Consequently, in this research, particle movement will primarily be used to separate PVs from other combinations. For instance, '*You must **put out** the* fire'[SI] can also be written as '*You must **put** the fire **out***' and *put out* is, therefore, regarded as a PV. On the other hand, '*Let's **talk about** the* situation'[SI] cannot be changed to *'*Let's **talk** the situation **about***', and *talk about* is thus not a PV, but rather a PrepV. If uncertainty should arise in identifying PVs, for which the particle movement test has no definitive answer, some of the tests mentioned previously might be applied with circumspection. Table 2.1 below supplies a summary of the tests that have been suggested for the identification of PVs.

Table 2.1 Tests used to identify PVs

Test type	Test design	Example	Notes
Unacceptable fronting in wh-questions	PV will not allow the particle to be moved before the verb in wh-question formation	*She **turned on** the stove.* **On what did she **turn**?* Not acceptable = PV	Does not separate all prepositions from particles (intransitive PVs included)
		*She **looked at** the view.* *At what did she **look**?* Acceptable = PrepV	
Unacceptable fronting in relative clauses	PV will not allow sentence to be turned into a relative clause	*She **turned on** the stove.* **The stove **on** which she **turned**…* Not acceptable = PV	Does not separate all prepositions from particles (intransitive PVs included)
		*She **looked at** the view.* *The view **at** which she **looked**…* Acceptable = PrepV	
Adverb insertion	PV will not allow adverb to be inserted between the verb and particle	*She **turned on** the stove.* **She **turned** quickly **on** the stove.* Not acceptable = PV	Does not separate all prepositions from particles (intransitive PVs included)
		*She **looked** favourably **at** the view* Acceptable = PrepV	

Particle placement	PV allows the particle to be placed before or after the direct object	*She **turned on** the stove > She **turned** the stove **on**.* Acceptable = PV *She **looked at** the view > *She looked the view at.* Not acceptable = PrepV	Separates prepositions from particles
Phrase fronting	PV does not allow the particle to move to the front of the sentence	***On** the stove she **turned**.* Not acceptable = PV ?***At** the view she **looked**.* Acceptable (though awkward) = PrepV	Does not separate all prepositions from particles (intransitive PVs included)
Passive voice formation	PV allows for the sentence to be passivized	*The stove was **turned on**.* Acceptable= PV **The bridge was **come across**.* Not acceptable = PrepV	Does not conclusively separate prepositions from particles
Verb substitution	PV can be replaced by a single word	*She **turned down** his offer > She rejected his offer.* Replaceable = PV *He **turned off** the road.* No replacement = PrepV	Does not conclusively separate prepositions from particles
Formation of action nominals	A noun can be derived from a transitive PV	*Her turning on of the stove…* Acceptable = PV **Her looking at of the view…* Not acceptable = PrepV	Separates prepositions from particles
Pronoun placement	PV requires the direct pronoun to be placed between the verb and the particle	*She **turned** it **on**.* Acceptable = PV **She looked it at.* Not acceptable = PrepV	Separates prepositions from particles
Sentence stress	Stress is required on both parts of the PV (or, at least, on particle)	*She **túrned on** the stove vs* *She **túrned ón** the stove.* Particle requires stress = PV *She **lóoked at** the view vs* *She **lóoked át** the víew.* Particle requires no stress = not PV	Does not conclusively separate prepositions from particles
Listing	PVs may simply be listed (not a true test)	For example, the PHaVE list (Garnier & Schmitt, 2015)	Successful use of list depends on definition used in compilation

2.4 Types of PVs

PVs display "a remarkable degree of semantic complexity" which necessitates categorisation according to degree of compositionality (Zarifi & Mukundan, 2014:52). PVs are, therefore, typically categorised as literal (the *verb proper + particle* combination retaining their original meanings), and idiomatic or figurative (the *verb proper + particle* combination producing a new meaning, unrelated to their separate meanings) (Darwin & Gray, 1999:68; Liao & Fukuya, 2004:196-197). According to Morales (2000a:7-8), literal PVs are considered "compositional" as their meanings can be deduced

from the individual meanings of their two parts, whereas idiomatic PVs are considered "non-compositional". Riguel (2014:112), in fact, uses "non-compositional" as an alternative term for the idiomatic category, as well as the term "opaque", and "directional" or "transparent" as alternative terminology for the "literal" category. Waibel (2007:6) views idiomaticity as a continuum, "with entirely opaque units at one end and entirely transparent ones at the other, the middle ground being covered by those where (at least) one element is transparent". Because of the obscurity of idiomatic PV meanings, these types of PVs seem to be particularly troublesome for L2 learners (Liu & Myers, 2020:407). On the other hand, while literal PVs are easier to understand, their patterns of use show that they have a limited range of uses.

Examples of literal PVs are *fall down* (Morales, 2000a:8), *take down* (Darwin & Gray, 1999:68), *go out*, *come in*, *take away* (Liao & Fukuya, 2004:196), and *stand up* (Bronshteyn & Gustafson, 2015:93). Yet several of these literal PVs could also be used idiomatically, as the following three examples illustrate. The PV *fall down*, apart from literally meaning *fall to the ground*, could mean *to fail*, as in '*This is where our plans* ***fell down***'[SI]**.** Similarly, *take down*, when used as a literal PV, means *to remove something from its fixed position*, but can also be used to mean *to overcome an enemy*, as in '*They* ***took down*** *the opposing team in an enthralling match*'[SI]. A further example is the PV *go out*, which means *to leave the structure you are presently in* when used in a literal sense, but which could mean *to commit to someone in a romantic way* if used in an idiomatic sense, as in '*They announced their engagement after* ***going out*** *for two years*'[SI].

Examples of idiomatic PVs are *turn down*, *pass out*, *run into* (Morales, 2000a:8), *make up* (Darwin & Gray, 1999:68), *let down*, *turn up* (Liao & Fukuya, 2004:197), and *butter up* (Bronshteyn & Gustafson, 2015:93). Nevertheless, some of the idiomatic PVs listed here could also belong to the literal category. For example, *turn down* means *to reject* when used idiomatically, but could also be used to indicate that something is being turned down physically, as in '*He* ***turned down*** *his collar*'. (This is also true of the PV *turn up*.) The PV *let down* means *to disappoint* when used in an idiomatic sense, but can also

be used in a literal sense. Consider, for example '***Let down*** *the car tyre'*. These examples confirm Waibel's (2007:16) statement that "many phrasal verbs exhibit a number of different meanings which can range from completely transparent to completely opaque". Consequently, it appears that the category to which a PV belongs is often dependent on the context in which it is used. Indeed, when the many meaning senses (Myers, 2018) of some PVs are taken into consideration, it seems apparent that classifying PVs without reference to the context is ineffective. An added complication is that PV meanings are also affected by the register in which they appear (Liu & Myers, 2020:406). This is, however, not the only issue that affects PV categorisation, as will be seen in the next paragraph.

Many researchers include a third category of PVs that is sometimes given as "completive" or "aspectual", in that it describes a completed action (Darwin & Gray, 1999:68; Liao & Fukuya, 2004:197; Waibel, 2007:19). Liao and Fukuya (2004:197) give *cut off* and *burn down* as examples of the particle describing the result of an action, while Darwin and Gray (1999:68) offer the PV *eat up*. Darwin and Gray (1999:68) describe aspectual PVs as having a *verb proper* with a literal meaning, whereas "the particle contributes meanings, [which may not always be] commonly understood, about the verb's aspect". As example, they provide "*They* ***ate up*** *all the chips and* ***drank up*** *all the soda*", where the particle *up* indicates that the actions have been completed. On the other hand, Morales (2000a:11) defines this category of PVs as "not transparent, but not idiomatic either", and provides *play around*, *think over*, and *mix up* as examples. Then, again, Bronshteyn and Gustafson (2015:93), with reference to *speak up* as an example, describe this category as depicting a verb which is literal and a particle which is not, suggesting that the particle drives the idiomaticity of this type of PV. Waibel (2007:19) adds that the aspectual particle, as in *eat up* and *burn down*, implies "entirely, completely" rather than "direction" or "movement from a lower to a higher position". Given the inconsistency of some of these definitions, some researchers prefer to leave the third category undefined. Hence, Gardner and Davies (2007:342) simply refer to it as "all degrees in between" the literal and figurative categories. Others, such as McArthur (1989:39), do not identify a third category at all.

To avoid this confusion, it may be argued that the best approach would be to classify PVs only according to the literal and idiomatic categories. This does not seem to solve the problem, however. As was seen from the discussion on defining the PV, there are researchers, most notably Bolinger (1971), who do not accept the literal category of PVs at all, and only use the aspectual (completive) and idiomatic (figurative) categories (McPartland-Fairman, 1989:42-43). Indeed, since the 18th century, some have considered only idiomatic PVs to be legitimate (Biber *et al.*, 1999; Thim, 2012:247). It seems that as much debate is aroused by classifying PVs as by defining them.

2.5 PV use across registers

As mentioned previously, it has traditionally been accepted that PVs are more prevalent in speech than in formal writing, and this viewpoint is still widely held today (Celce-Murcia & Larsen-Freeman, 1983:265; Chen, 2013a:426; McArthur, 1989:39,40; Myers, 2018:11). For example, Chen (2013a:426) maintains that avoiding the PV in academic writing is in line with the writing conventions of academic institutions. Furthermore, Siyanova and Schmitt's (2007:121) contention that the real challenge for L2 learners is "choosing the verb which has the appropriate register, and *which conforms to the expectations of the speech community*" rather than "choosing the verb form *which carries the correct meaning*" [my emphasis], indicates the far-reaching effect that this belief has had on PV use. Zhou and Wang (2024:2), while agreeing that PVs are unsuitable for use in academic English, expand the debate somewhat by asserting the usefulness of PV use in business English "for effective communication".

The notion of the informality of PVs is apparently linked to their history, being considered a "plebian" word form used in Old English (± 450 AD to ± 1100 AD) and in Middle English (±1100 AD to ±1500 AD), and then superseded by Romantic (Latinate) word forms which were preferred in all formal writing (McPartland-Fairman, 1989:26-27) (see §2.2). Indeed, Kennedy (1920:33), whose work on the PV is considered seminal and whose influence on this topic has consequently been significant (Dixon, 1982; Gorlach, 2008; Kovács, 2002; Thim, 2012; Wild, 2010), regards the choice of the PV over its Latinate one-word synonym as a clear sign of being improperly schooled, as "the few who take a real pride in

the precision and dignity of their English" would avoid PV use. Furthermore, Kennedy (1920:37) bemoans the fact that "a large number of simple verbs of more highly specialized meaning are being crowded out of general use". Therefore, it is significant that Kennedy (1920:41) should acknowledge that many figurative PVs are richer and more descriptive than their more formal one-word alternatives, and that he finds even some literal PVs to be more expressive than the "impersonal" single-verb choices that are available. However, he also warns against the arbitrary use of PVs when a one-word verb with a more exact meaning exists (Kennedy, 1920:45).

In contrast to the views expressed above, there are several researchers who do not support the idea of the PV being primarily colloquial. Thim (2012:122) derides the notion of the PV being unworthy of formal use as "the avoidance of phrasal verbs by pedantic speakers". To substantiate his argument, Thim (2012) points out that 18th century dictionaries show no aversion to PV use, and he further claims that the notion of the PV being "distinctly colloquial" is rather a result of the "well-known objection to stranded prepositions, with a hypercorrection banning all particles (and monosyllables) from sentence-final position". He asserts that some of the examples (such as *chip in*, *cough up*, *hang out* and *knock off*) used by Kennedy (1920:34) to prove his point about PVs being colloquial are undeniably slang, but that this does not mean that all PVs should be avoided (Thim, 2012:122). Indeed, Kennedy (1920:34) does not seem to condemn the use of all PVs either. He lists only 22 specific PVs (out of the 826 PVs that formed part of his research), the use of which he feels distinguishes between people of good and minimal education.

Thim (2012:122) also refutes the assumption that the existence of a PV synonym automatically presumes PV colloquiality, stating that some suggested synonyms "can only be regarded as 'synonyms' by a very wide application of the term". Garnier and Schmitt's (2015:664) contention that PVs "carry connotations that their single word equivalents do not have" supports this point. Apart from a weakening of the intended meaning, simply swapping one for the other can lead to language use that appears awkward. For example, in the sentence "*They jumped into the car and **took off***", the sense of

leaving or departing, especially suddenly, is not quite conveyed if *took off* is replaced by *leave* or *depart* (‘*They jumped into the car and **left**’*[SI]). Similarly, one meaning of the PV *work out* is *happen or develop in a particular way*, normally used in conjunction with the adverbs *well* or *badly*, as in the sentence *“Everything **worked out** well in the end”*. In this case, replacing *work out* with *happen* or *develop* not only weakens the meaning, but also creates language that is awkward (?‘*Everything happened well in the end*’[SI]) (examples taken from Garnier and Schmitt (2015:658), unless otherwise indicated). Furthermore, while the *Collins COBUILD Phrasal Verbs Dictionary* (2013:514) states that, “[i]n English, it is quite common for a phrasal verb to have a single-word equivalent”, according to Thim (2012:122), a large number of PVs have no synonyms. Even Kennedy (1920:34), whose preference for one-word Latinate alternatives (what he refers to as “more highly specialized verbs”) to PVs was mentioned in a previous paragraph, could find synonyms for only 110 of the 826 PVs that he examined. Furthermore, he admits that PVs, even those with synonyms, “are possessed of a variety of meanings”, which suggests that the elimination of any PVs from the vocabulary would be detrimental to the language (Kennedy, 1920:34). Moreover, synonyms constitute a general feature of the English language, and are not particular to PVs (Thim, 2012:122).

As further proof of the general applicability of PVs, several researchers maintain that PV use is evident across a range of registers (Alangari, Jaworska & Laws, 2020:1; Darwin & Gray, 1999:66; Liu & Myers, 2020:407; McPartland-Fairman, 1989:1). Alangari *et al.* (2020:10) report a high frequency of PV use in expert academic writing in Linguistics. Darwin and Gray (1999:66) cite the Gettysburg address (*brought forth*) and the King James version of the Bible (*lie down*) as instances of eminent documents that make use of PVs. These examples illustrate the fact that “the phrasal verb is virtually unavoidable without lengthy and often pretentious circumlocutions” (Darwin & Gray, 1999:66).

A closer look at actual PV use, such as the extensive research done by Biber *et al.* (1999), might help to clarify the argument. Biber *et al.* (1999) looked at PV use across four registers, namely speech (conversation), fiction, news and academic discourse. It should be noted that some researchers, such

as Liu and Myers (2020:407), limit their investigation to speech and academic discourse only, as these "arguably occupy the two ends of the language formality continuum". However, Biber *et al.* (1999), in considering all four registers, allow us to obtain as complete an idea as possible of PV use across registers. They found PV use (especially that of the intransitive PV which often functions as an imperative, e.g. *hold on* and *shut up*) to be principally used in fiction and speech (Biber *et al.*, 1999:408). Thus, the most common PV, namely *come on,* is an intransitive PV which is encountered predominantly in conversation and fiction, hardly ever in newsprint, and almost never in academic discourse. On the other hand, the PV *go on*, another very commonly used PV, appears equally across all registers (Biber *et al.*, 1999:411). For example, one will see "***Go on**. Stamp on it*" in conversation, "*If it failed once, there's no point in **going on***" in fiction, "*Labour must **go on** getting the public's support by constructing strong unity of purpose*" in the news, and "*As time **went on**, Liebig developed his thesis*" in academic discourse (Biber *et al.*, 1999:411). Biber *et al.* (1999:412) note that, generally, transitive PVs are found more frequently across all registers than are intransitive PVs, and some, such as *carry out, take up, take on, set up*, and *point out,* are more prevalent in writing than in speech. Liu (2011:678) adds *bring about* to this list, and also notes the high frequency of the PVs *break down, carry on, follow up, make up, rule out*, and *sum up* in both speech and academic writing. In fact, in Liu's study (2011), the 150 PVs that were identified as the most frequently used all appeared in academic writing, and PVs such as *carry out* and *point out* appeared to be particularly useful in that register (Liu, 2011:680). Hence, it would seem that PVs are indeed present in all registers, although not all PVs are present in all registers. (Interestingly, Biber *et al.* (1999) used the British National Corpus (BNC) in their research, whereas Liu and Myers (2020) and Liu (2011) used the Corpus of Contemporary American English (COCA), and the similarity in PV use across British and American English is noteworthy and informative.) Liu and Myers (2020:407) make a further claim regarding PV use across registers, namely that the meaning senses of a PV have a significant impact on the register in which the PV functions. This aspect will be discussed in more detail in the next section.

The above discussion about PV use across registers leads to the following conclusion: the idea that PV use should be confined to informal writing and speech does not seem to be borne out by research. On the contrary, as mentioned above, PVs are to be found in all registers, although not all PVs function equally in all registers. This means that there are indeed some PVs that should be used only in speech and informal writing (for example, *chill out*), and, as with all other English expressions that would be considered "slang", should be avoided in academic writing. In this regard, Kennedy (1920:10) makes the interesting observation that, although some PVs should be considered colloquial, "only a year will suffice, sometimes, to transfer one from the lower stratum of linguistic society to a place of prominence and good standing". As example, he cites the PV *carry on,* which gained popularity, and respectability, during the first world war.

Another conclusion that can be drawn from the previous discussions is that the existence of a synonym for a PV does not automatically equate to its usage in preference to that of the PV, but, rather, that the context of the sentence should prescribe which word form contains the correct connotation. These conclusions strongly refute the notion that PVs are colloquial and suitable only for use in everyday communication and writing.

These differing opinions on the use of PVs in academic writing might be summarised as follows. While there is adequate proof that PVs have a place in academic writing, and that their strict avoidance might lead to awkward and unnecessarily verbose phrasing, they should be used with discretion. Of course, PVs that fall within the category of informality or even slang should be avoided at all cost.

2.6 Frequency of PV use

Current awareness of the importance of the PV is evident in the relative abundance of dictionaries focussed on identifying and defining all instances of this grammatical phenomenon. Examples of such dictionaries are the *Oxford Phrasal Verb Dictionary*, *Cambridge Phrasal Verb Dictionary*, *McGraw-Hill's Essential Phrasal Verbs Dictionary*, *Longman Phrasal Verbs Dictionary*, and *Collins COBUILD Phrasal Verbs Dictionary*. In trying to list every known PV, usually along with its various meaning senses, these

dictionaries are of immense value. However, for teaching purposes, Gardner and Davies (2007:343) query the efficacy of the dictionaries. Rather, they are of the opinion that the identification of high-frequency PVs would be more effective than listing every PV if the vast and ever-increasing number of PVs is taken into consideration. Furthermore, it is very probable that such a vast amount of information could overwhelm L2 learners (Garnier & Schmitt, 2015:659). The emergence of multi-million-word corpora has made the tracking of high-frequency PVs possible, counteracting the arbitrary nature of a great deal of PV learning material, where the choice of which PVs to teach has often been based on intuition (Darwin & Gray, 1999:66; Schmitt & Redwood, 2011:175). Indeed, Trebits (2009:471) highlights "the relevance of the use of specially designed corpora built by language professionals to suit the specific needs of their students and / or research purposes in developing teaching materials".

The importance of reporting on high-frequency PVs is supported by the fact that, despite their large number, some PVs appear infrequently in the language, whereas "a small number of the most frequent PVs account for the majority of the uses of PVs in English" (Liu & Myers, 2020:405). Learning material based on high-frequency PVs will consequently be more reflective of "authentic language use" (Garnier & Schmitt, 2015:649). Moreover, research suggests that L2 learners produce the high-frequency PVs that they encounter in their learning materials more often than low-frequency PVs (Chen, 2013a:421).

Measuring frequency within meaning senses has also been recognised as important, because of the polysemous nature of many PVs (Garnier & Schmitt, 2015:648; Liu & Myers, 2020:405). As illustration, Liu and Myers (2020:405) refer to the PV *make up*, which has several distinct meaning senses, such as *compensate, compose*, *fabricate* and *put on cosmetics*. Kennedy (1920:35) also mentions 15 meaning senses for the PVs *put out*, *take up* and *set up*, 12 for *get up*, and 11 for *take in* and *turn out*. On the other hand, not all meaning senses of a PV are necessarily used to the same extent, and, accordingly, not all meaning senses of a PV would form part of a list of high-frequency PVs. Furthermore, Alangari

et al. (2020:9) point out that "when used in academic writing, PVs have different senses from those used in spoken or more general language, many of which are the less frequent ones".

Gardner and Davies (2007:346) support the idea of corpus-based research including frequency of meaning senses, and not frequency of word forms only, to avoid the oversimplification that results when no differentiation is made between these two aspects. They view this as "one of the fundamental issues to be addressed as corpus linguists attempt to build bridges to language education" (Gardner & Davies, 2007:346). It is worth noting that, while PV dictionaries are not deemed useful in pointing out frequently used PVs, Liu and Myers (2020:409-410) consider them valuable sources of finding the various meaning senses of a PV. On the other hand, Garnier and Schmitt (2015:659) caution that there are many inconsistencies to be found among the different dictionaries as far as meaning senses are concerned. Consequently, to obtain reliable results, a variety of PV dictionaries should be referenced.

The research conducted by various researchers into high-frequency PVs, and, in some cases, including PV meaning senses, will be discussed below. As no prior research has been done on PVs in South Africa, comparisons are only possible with international studies.

2.6.1 Gardner and Davies (2007)

The viewpoint that Gardner and Davies (2007) took in investigating PV frequency, using the BNC, was to identify the verbs that are most frequently used in PVs, as well as the most frequently used particles. In identifying these elements, Gardner and Davies (2007) hoped to ease the learning burden for the L2 learner. For example, the fact that o*ut, up, down* and *back* are more likely to be used as particles than as prepositions, in contrast to *under*, *by* and *across* which are hardly ever used as particles, could help learners to identify PVs more easily. Furthermore, they found that 20 lexical verbs appeared consistently in 53.7% of the PVs in the corpus, which indicated the prevalence of certain verbs in PV constructions (Gardner & Davies, 2007:348). Ten of these verbs also featured on the list of the most prolific lexical verbs in the BNC. These 20 verbs combined with a limited number of particles (*out, up,*

on, back, down, in, over, and *off*) to form about half of the PVs in the BNC (50.4%), although not every *verb-adverbial particle* combination was possible (Gardner & Davies, 2007:348,350). Interestingly, the verbs *pick*, *point*, and *carry* (70%, 52% and 51.1% respectively) were found to appear more often as PVs than as single lexical verbs (Gardner & Davies, 2007:348).

This research resulted in a list of the 100 most frequently used PVs, representing 51.4% of the PVs in the BNC corpus. A staggering 12 408 PVs make up the other 48.6% of PVs in the corpus, which clearly indicates that teaching L2 learners about the 100 most frequently used PVs will provide them with the ability to manage the PVs they are most likely to encounter. When the various meaning senses of the 100 PVs are included, the list increases to 559 entries, which the researchers still believe to be manageable for L2 learners. Furthermore, considering that a PV like *break up* has 19 different meaning senses (the average being 5.6 meaning senses per PV), one can surmise that not all of the meanings of a PV are likely to occur equally frequently, and that they do not all need to be covered in L2 learning material. Garnier and Schmitt (2015:654) make the interesting observation that the high number of meaning senses per PV "means that, in reality, the learning load of PVs is probably greater than most other words or word combinations in English".

It should be noted that Liu and Myers (2020:405) found the Gardner and Davies (2007) study limited since it only investigated PVs of which the lexical verb fell within a list of the 20 most frequently used verbs, possibly resulting in many useful PVs being ignored. They also found the absence of information on PV use frequency across registers to be a shortcoming. Garnier and Schmitt (2015:655-657), moreover, criticised the Gardner and Davies (2007) study because of inconsistencies in the meaning senses: in some cases, the meanings overlapped, were redundant, or insufficient. For these reasons, Garnier and Schmitt (2015:657) felt that meaning senses should not be derived from only a single data source such as Wordnet, the tool used in the Gardner and Davies (2007) study.

2.6.2 Liu (2011)

In a corpus-based study, Liu (2011) compared PV use in American and British English, and across registers in the case of American English, with the aim of identifying the most commonly used PVs, as well as providing information on patterns of PV use. The study was partially based on the research previously conducted by Gardner and Davies (2007). For this reason, Liu (2011:663) used the same definition of a PV as Gardner and Davies (2007), namely that of a verb proper, followed either contiguously or non-contiguously by an adverbial particle. Liu (2011:662), however, attempted a more in-depth study, and took the acknowledged shortcomings of the previous study into account, such as their study making use of only the top 20 lexical verbs that produce PVs, and, therefore, not including any PVs using verbs not on that list. Secondly, as Gardner and Davies (2007) had made use of only one corpus, namely the BNC, their results were confined to British English only. Thirdly, they did not include cross-register PV data, which means that the context in which PV meanings were used was not made available.

In this study, the COCA was primarily used, with the BNC also providing data both directly and indirectly. Liu (2011:665) used the COCA interface search functions, firstly, to find all PV combinations for specific lexical verbs (e.g. *go + particle*); secondly, to find all PV combinations where one word separates the lexical verb and particle (e.g. '*She **took** him **down***'[SI]); thirdly, to find all PV combinations where two words separated the lexical verb and particle (e.g. '*She **took** her adversary **down***'[SI]); and, finally, to capture the results per register in an Excel sheet.

While confirming much of Gardner and Davies' (2007) research, Liu's (2011) study resulted in a more comprehensive list: 150 of the most frequently encountered PVs in both British and American English were included, and more information was provided about PV use than given in the Gardner and Davies (2007) study, such as the main register in which each PV operates. The information on the list was designed to be accessed according to the intended purpose: by variety of English (British or American),

or by register (formal or informal) (Liu, 2011:679). Interestingly, Liu (2011:679) found little difference between the American and British frequency lists.

The research resulted in a list of the 150 most frequently used PVs which can be accessed according to the required learning purpose: either the American or British English PV frequency list, or a particular register, may be selected (Liu, 2011:679). Liu (2011:679-680) also points out several ways in which the information can be used effectively by teachers. For example, while PVs are predominantly thought to be used in an informal sense, some PVs, such as *carry out* and *point out*, often appear in academic writing. Furthermore, while literal PVs are easier to understand, their usage patterns show that they have limited functions, which suggests that it is far more beneficial to concentrate on learning idiomatic PVs.

The significance of register is moreover emphasised by the fact that specific meanings of polysemous PVs are attached to specific registers. This is illustrated by the PV *make up* for which Liu (2011:676) identified four meanings: *compose* or *constitute* (as in '*Men used to* ***make up*** *the labour force*'[SI]), *decide* (as in '*She* ***made up*** *her mind to go*'[SI]) , *compensate* (as in '*She* ***made up*** *for her lack of skill by working hard*'[SI]), and *fabricate* (as in '*They* ***made up*** *a story to tell their parents*'[SI]). Liu (2011:676) investigated the first hundred tokens of *make up* in the spoken and academic registers of the COCA, and found that meaning differed widely across the two registers. The four meanings were relatively evenly represented in the spoken English corpus, with *compose* at 12%, *decide* at 27%, *compensate* at 26%, and *fabricate* at 25%. On the other hand, in the academic register, *compose* was the most represented meaning by far with 79%, followed by *compensate* at 18%, *fabricate* at 2%, and *decide* at 1%. Interestingly, Liu (2011:679) notes that there is little difference between the American and British most frequently used PVs.

2.6.3 Garnier and Schmitt (2015)

In response to Gardner and Davies (2007) and Liu (2011), Garnier and Schmitt (2015:648) hoped to improve the usefulness of the resultant list by the inclusion of the most frequently-used meaning

senses of high-frequency PVs. (While Liu (2011) did acknowledge the importance of taking a PV's meaning senses into consideration, lack of space prevented such an inclusion in his study.) As Garnier and Schmitt (2015:648) put it, "[j]ust as priority should be given to the PVs that occur most frequently in language, priority should also be given to those meaning senses that occur most frequently for any individual PV." The inclusion of only the most frequently used meaning senses aligned with the intended practicality of this study as too long a list would not be useful to either learners or teachers (Garnier & Schmitt, 2015:662). They based their research on the list created by Liu (2011) as they deemed the process used in the creation of that list to be credible. Thus, they considered the items on the list to correctly represent the most frequently-used PVs. This resulted in the creation of the **PH**rasal **VE**rb Pedagogical List (the PHaVE List), which includes, for each high-frequency PV, the most relevant meaning senses, the percentage of frequency of each meaning sense, and an example sentence. The total number of entries (with each meaning sense considered an entry) is 288, which is substantially fewer than previous lists. All registers were represented equally to ensure that the information was as widely applicable as possible (Garnier & Schmitt, 2015). Although this list seems vastly more manageable than earlier ones, the researchers suggest, moreover, that semantic transparency be taken into consideration, as literal PVs will take less time to teach than figurative PVs.

2.6.4 Myers (2018)

In his 2018 research, in a study similar to that of Liu (2011), Myers (2018) likewise set out to identify the 150 most frequently used PVs, also using the COCA. However, this study aimed at including the meaning senses of PVs across registers. Furthermore, in contrast to the Liu (2011) study, it did not include a British corpus. The aim of the research was to address the problem that polysemous PVs pose for L2 students by providing a list of prominent PVs and their various meaning senses, thus expanding on the PHaVE List created by Garnier and Schmitt (2015).

Though the COCA comprises five subcorpora, only the speech and academic writing subcorpora were used in the study, as, firstly, these two subcorpora "may be considered to be at the two ends on a

style or register continuum", which means that use and meaning variations will be most apparent between these subcorpora (Myers, 2018:7-8). The second reason was that these are the two registers L2 students are most likely to use (Myers, 2018:8).

Using the 150 most frequently used PVs identified by Garnier and Schmitt (2015), Myers (2018) first assembled a catalogue of all meaning senses ascribed to each PV, using various academic sources, including Garnier and Schmitt's (2015) PHaVE List. He then extracted 600 tokens of each PV from the speech subcorpus, and 600 tokens of the same PV from the academic writing subcorpus. Thereafter, the appropriate meaning sense was identified for each token, using the catalogue of all meaning senses previous compiled. This resulted in a list that indicated not only the most used meaning senses per PV, but also included the register in which those meaning senses appeared.

It is interesting to note that the meanings of the PVs were not always clear even to the researcher, which underscores the probability of L2 students encountering issues with their understanding of PVs (Myers, 2018:10). Meaning senses proved to vary markedly across these registers, which meant that context became important to the correct understanding of the way in which a PV was used in a particular register (Myers, 2018:27). In fact, Myers (2018) considers the subjectivity required to interpret all the meaning senses of a PV as one of the limitations of the study.

As many as ten meaning senses were found for some of the frequently used PVs, spread across the two registers. The following examples are those identified by Myers (2018:37-38). The PV *go out* has four meaning senses belonging to the spoken word register (*to go on a date or to a specific location* (46.4%); *to exit or leave, to move out or depart, to send out or announce* (42.5%); *to take the field or go on a mission, oftentimes with a specific goal in mind* (10.1%); *to extinguish or be extinguished, to be cut off or eliminated* (6.8%)), and five belonging to the academic discourse register (*to exit or leave, to move out or depart, to send out or announce* (35%); *to go on a date or to a specific location* (28.4%); *to extinguish or be extinguished, to be cut off or eliminated* (14.9%); *to do something extra for someone - go out of someone's way* (idiom) (10.2%)) (Myers, 2018:37-38).

The vast variety of possible meanings show the importance of taking not only the meaning senses of PVs into account, but also the registers in which the various meanings are most used. For instance, if L2 learners were to learn high-frequency PVs regardless of register, they could very well become confused when encountering a familiar PV in an unfamiliar register, and with an unfamiliar meaning.

The researcher comments that it might also be useful if L2 learners were taught the meaning senses they are most likely to encounter. (This aspect was, of course, the basis of Garnier and Schmitt's (2015) research, which resulted in the *PHaVE List* (see above)). As deductive teaching of all the most important PVs, including their most frequently used meaning senses across the two registers, would be a cumbersome task for teachers, Myers (2018:31) suggests using an inductive, corpus-based teaching approach. For instance, the teacher could print out some PV tokens from the COCA, and allow L2 learners to deduce the meanings of the PVs. Afterwards, their responses could be compared to Myers' (2018) own to elicit discussion and further exploration of the COCA. Consequently, L2 learners would, hopefully, develop the ability to discover the correct meaning sense of a particular PV token.

2.6.5 Liu and Myers (2020)

In a similar study, Liu and Myers (2020) explored the main meaning senses in both speech and academic discourse of the 150 most frequently used PVs, as identified by Liu (2011). This was a follow-up study to and extension of Garnier and Schmitt's (2015) research. Several aspects of their research are, thus, aligned to that of Garnier and Schmitt (2015). For example, in the list that was subsequently produced (named the S&A PHaVE List, clearly indicating its connection to Garnier and Schmitt's (2015) list, but with the inclusion of Spoken and Academic Writing (S&A W) in the name), only the PV meaning senses encountered a significant number of times were captured. This was in line with the researchers' purpose of listing only those entries that would most repay the time spent learning them. Furthermore, as in the Garnier and Schmitt (2015) list, they included example sentences (Liu & Myers, 2020:410). However, a distinction of this research was the identification of three separate groups of PVs, which indicated usage patterns across the two registers: in the first group, to which 48% of the

150 PVs belonged, there was a substantial difference in the meaning sense and PV frequency across the two registers (examples are *break up*, *make up*, and *come up*); in the second group, to which 22.66% of the PVs belonged, there was no substantial difference in the meaning senses across the two registers, but there was a substantial difference in PV frequency (examples are *come back* and *find out*); and, in the third group, to which the remaining 29.34% belonged, there was no substantial difference in either meaning sense or frequency across the registers (Liu & Myers, 2020:413). This research, consequently, demonstrated the importance of register in PV use (Liu & Myers, 2020:433).

2.6.6 Schmitt and Redwood (2011)

Schmitt and Redwood (2011) investigated a novel aspect of high-frequency PVs in their research, focusing on L2 learner knowledge of these PVs, and the means by which this knowledge was gained. In other words, did L2 learners, in fact, know high-frequency PVs? Based in England, this study required sixty-eight English students (both of English as a second language and English as a foreign language) to complete two tests based primarily on the high-frequency PVs identified by previous researchers (Gardner & Davies, 2007), with the addition of a few less frequently-used PVs. One of the tests assessed learners' ability to produce PVs (productive skills), and the second test assessed learners' ability to recognise PVs (receptive skills). Learners' biodata were also obtained by means of questionnaires. The test results indicated that PV frequency was a good predictor of productive knowledge, but not of receptive mastery. Thus, while it was evident that PV frequency did play a role in L2 PV knowledge, this factor did not on its own explain the variations found in the test results. Rather, other factors (gleaned from learners' biodata) were found to be more relevant in PV acquisition, namely L2 learners' English proficiency, non-academic reading, and exposure to English television and films (Schmitt & Redwood, 2011:189). The researchers concluded that "phrasal verbs are idiosyncratic in terms of learning burden, and that a purely frequency-based explanation can never fully explain their acquisition" (Schmitt & Redwood, 2011:187).

In summary, then, it is clear that the identification of the most frequently used PVs is important in preventing the L2 learner from being overwhelmed by the vast number of PVs that would otherwise have to be learned. However, frequencies should not be considered in isolation, but should take the meaning senses of PVs into account, as well as the registers to which the various meaning senses belong.

2.7 Main issues regarding the use of PVs

Previous research has identified various areas of concern with regard to the use of PVs by L2 students in particular, such as semantic transformation, the polysemous nature of PVs, the obscurity of idiomatic PVs, and the influence of a mother tongue in which a PV structure is absent. These concerns are discussed in this section. Non-standard PV use by L2 students will also be considered.

2.7.1 Semantic transformation

It has been suggested that the structure of the PV, being a two-word verb that functions as a single lexical unit, creates difficulties for the L2 learner. Unable to recognise the unity of the PV, learners instead derive meaning by applying the original, separate meaning of each word to the two-word combination (Garnier & Schmitt, 2015:646; Siyanova & Schmitt, 2007:119). While using this process to find the meaning of a PV might not always result in error (as in the case of the literal use of *lie down,* in '*Why don't you **lie down** and have a rest*'[SI]), there are instances where inaccuracy is highly likely (as in the case of the idiomatic use of *lie down*, in '*We're not going to **lie down** and accept this abuse*'[SI]). For instance, Schmitt and Redwood (2011:174) argue that:

> the problem for learners is that these frequent and apparently simple components may come together to form units which are specialised, emotive, and idiomatic (e.g., *the situation is really getting her down; I can't make out what this says; don't give up now; it was too much to take in*).

The solution, according to Siyanova and Schmitt (2007:119), is for PVs "to be acquired, stored and retrieved from memory as a holistic unit". Bronshteyn and Gustafson (2015:92) offer a variation of

this solution, as they believe that the different categories of PV are managed differently from a cognitive perspective. For example, they maintain that literal PVs, by the very fact of their transparency, are stored as two separate units "in the mental lexicon", while more idiomatic PVs are stored as single units. In the light of this theory, they suggest that PVs should be taught according to their categories: literal PVs taken at face value, figurative PVs memorised, and aspectual PVs as having (according to their own definition) a literal verb proper but a figurative particle (Bronshteyn & Gustafson, 2015:94). However, Haugh and Takeuchi (2023:668) are not in favour of the memorisation of PVs as a solution, as the superficial PV knowledge of their Japanese L2 students was a result of having to "remember PVs as unanalyzed [sic] chunks from wordlists". Rather, some researchers (Badem & Simsek, 2021:55; Wei, 2021:35; Zhou & Wang, 2024:19) suggest incorporating contextual and authentic L2 input in the teaching of PVs to avoid mechanical PV memorisation. The lack of research into PVs in the South African context has meant that no recommendations have been made as to how they should be taught in SA classrooms.

The issue of semantic transformation is more complex than the supposed transparency or opacity of some types of PVs. The role of the particle can certainly be said to add complexity to an understanding of the PV. McPartland-Fairman (1989:41-42) draws attention to the essential, though diverse, roles that particles play in adding nuanced meaning to the PV, at times modifying (for example, *jot down*) or intensifying (for example, *clear out*) the verb proper. White (2012:420), moreover, highlights research that illustrates how some particles (in this case *down* and *out*) "systematically contribute meaning to phrasal verbs". The particle appears to be of importance even in cases where it does not have an obvious role, as, according to McPartland-Fairman (1989:41), "it seems that the verb can't fully express what the combination does even if the particle is redundant". McPartland-Fairman (1989:41) gives as an example the PV *fill up, where, "*initially, the particle was added to give emphasis or to round out the speech rhythm". Haugh and Takeuchi (2023) confirm the importance of making L2 learners aware of the particular role of adverbial particles in PV combinations, since "simply teaching learners this fact could alleviate the deceptive transparency projected by PVs".

Kennedy (1920:9) distinguishes between three different uses of the particle. In the first instance, the particle and the verb proper give up their meanings to form a compound with a new meaning (for example, *own up*, as in *confess*). Secondly, a "weak" particle only slightly modifies the verb proper (for example, *blot out*, as in *destroy*). In this case the particle loses its normal adverbial function, yet it "very seldom loses its adverbial personality so completely that it makes no difference in the use of the combination" (Kennedy, 1920:9, 37). Some of the examples provided by Kennedy (1920:28) illustrate this point: while the particles in *fall down* and *rise up* seem to add no value (as, in falling, one can only go down, and, in rising, one can only go up), some meaning does seem to be lost if these particles are discarded. For example, there is less sense of the degradation of the fall without the particle *down* added in *fall down*; similarly, *rose up* has a greater sense of collaborative or joint protest with the addition of the particle *up*. Compare, for example, '*Upon entering the presence of the king, they* ***bowed down***'[SI] to '*Upon entering the presence of the king, they* ***bowed***'[SI], and '*The people* ***rose up*** *in revolt*'[SI] to '*The people* ***rose*** *in revolt*'[SI]. The third particle use that Kennedy (1920:9) refers to is where both the particle and verb proper retain their original meanings (for example, *fall down*, as in *collapse*). These three *verb proper* + *particle* combinations, at first glance, may seem to correspond with figurative, aspectual and literal PVs, respectively. However, a closer inspection of the examples Kennedy (1920:9) provides suggests that this is not necessarily the case. For example, *burn down* is given as an example of each element retaining its meaning (the third case mentioned above), yet it is, in fact, an aspectual PV. In light of these complexities, Chen's (2013a:420) reference to this feature of the PV as "semantically non-computational" seems to have some validity.

2.7.2 Syntactic transformation

Apart from semantic transformation, the syntactic changes arising from the combination of a verb proper and an adverbial particle may cause additional issues for the L2 learner (McPartland-Fairman, 1989:40-41). McPartland-Fairman (1989:40-41) supplies the following examples: a change from intransitive to transitive verb (*cough* and *cough up),* from transitive to intransitive verb (*clear* and *clear*

up); new verb formation, in which the verb proper in the *verb + adverbial* formation does not exist separately as a verb (*gun down, but not *gun)*, or the verb proper has a vastly different meaning to its meaning as a PV (*cracked* and *cracked up*).

McPartland-Fairman (1989:41) further points out the possible confusion created by inconsistent *verb + particle* combinations, with some verbs combining with a range of particles (for example *get* ***on***, *get* ***up***, *get* ***out***, *get* ***in***, *get* ***through***, *get* ***over***, *get* ***down***, *get* ***ahead***), while others combine with a limited number of particles (for example *set* ***up***, *set* ***out***, *set* ***off*** and *set* ***down***). In some instances, the verb combines with only one particle (for example ***freak out***). These syntactic changes could further hamper PV acquisition because of the added grammatical load faced by the L2 learner.

2.7.3 Syntactic flexibility

The ability of the particle to be separated from the verb proper is an essential feature of the PV, as is evident from the inclusion of this feature in the definition, where the particle is described as being "either contiguous (adjacent) to [the] verb or noncontiguous (i.e., separated by one or more intervening words)" (Gardner & Davies, 2007). Such syntactic flexibility refers primarily to transitive PVs, where the direct object may be inserted between the verb proper and the particle (Darwin & Gray, 1999:69; White, 2012:420). This feature seems to counteract the concept of the syntactic unity of the two-word PV, which, as discussed previously, already poses a problem for the L2 learner (Garnier & Schmitt, 2015:646; Siyanova & Schmitt, 2007:119). Furthermore, as this separability is a feature that is unique to the PV, and is not found in other English word groups (Chen, 2013a:420), learners will have no reference point for dealing with it. White (2012:420), therefore, sees this peculiarity of the PV as another one of the "hurdles" that learners have to face.

2.7.4 Polysemy

A further difficulty that PV use presents, especially in the case of figurative and aspectual PVs, is their polysemy (Chen, 2013a:420; Myers, 2018:4; Schmitt & Redwood, 2011:174; Siyanova & Schmitt,

2007:120; White, 2012:419). It has been estimated that there are at least 3 000 PVs in the English language, and, by some estimates, as many as 5 000 (Alhatmi, 2023:4). Garnier and Schmitt (2016:31) and Zhang and Ju (2019:1077) surmise that, as a consequence of the continual use of some PVs, a variety of idiomatic meanings evolve from their original literal meanings. This results in some PVs having a range of meanings with which students have to contend (Chen, 2013a:420; Garnier & Schmitt, 2016:31; Lu & Sun, 2017; Mahpeykar & Tyler, 2015:2). In fact, Gardner and Davies (2007:353) have pointed out that approximately 5.6 meaning senses can be contributed to each of the 100 most frequently used PVs. For example, six meanings have been attributed to the PV *take up*: *to discuss or deal with an issue, idea, matter, etc*.; *to use a particular amount of time or effort*; *to begin a specific job, activity, or hobby*; *to grasp an object, often moving it from a lower to a higher position*; *to assume a position, especially in relation to something else*; and, *to absorb or allow water, chemicals, to seep inside something*. The PV *get down* has seven meanings attributed to it: *begin to pay serious attention to/complete something or really delve into a plan*; *lower one's body as by kneeling, sitting or lying*; *come down from something*; *descend (car, horse, tree, etc.)*, *to move to a different location laterally, or relating to moving to a different location that is not necessarily relating to a movement from a higher to a lower place*; *to decrease or to reduce price, time, scores, etc*.; *to center or focus on one specific element, or to boil down to a certain point*; and, *to temporarily let yourself act in an unnatural way either to party or to complete a difficult job*. (These word senses are taken from Myers (2018)). Particular PVs, such as *break up,* might have up to 19 different meaning senses (Gardner & Davies, 2007:353). White (2012:420) reports that 21 meaning senses are given for the PV *go on* in *The American Heritage Dictionary of Phrasal Verbs* (2005).

The problem posed by this issue is the burden it places on L2 learners, who are expected to be familiar with the various meaning senses. Garnier and Schmitt (2016:37), in their research into student knowledge of the multiple meaning senses of PVs, found that L2 learners (in this case, first- to fourth-year BA students at two Chilean universities) displayed awareness of "only about 40% of PV meaning senses on average, with only about a 20% chance that all the various meaning senses of each PV tested

would be known". While the researchers did not establish the English proficiency levels of these students prior to investigating their knowledge of PVs, it is not unreasonable to expect at least moderate proficiency at their level of study. To add to the problem that learners face, "the multiple meanings of these lexical units have been extensively regarded as arbitrary and unpredictable" (Mahpeykar & Tyler, 2015:2).

Schmitt and Redwood (2011:174) suggest that PV polysemy is not insurmountable for the L2 learner because "a semantic link may be made between the different senses (cf. *fill in a hole, fill in a form, fill in somebody on something, fill in for somebody*)". On the other hand, White (2012) argues that such semantic relationships are often far from clear. In fact, Schmitt and Redwood (2011:174) admit that polysemy has degrees of transparency, as, in some cases, "the connection is more tenuous (cf. *put up a fence, put up a fight, put somebody up for the night*), and the meanings more difficult to interpret" [author's italics]. Liu (2011:680) further points out the added complication of the various meaning senses being situated across registers.

Recently, through the use of Cognitive Linguistics, some attempts have been made to address this difficulty (Haugh & Takeuchi, 2023; Mahpeykar & Tyler, 2015; Yasuda, 2010). For example, Yasuda (2010:251) maintains that particles are "orientational metaphors" linked to the spatial orientation that human beings experience, for example *up* and *down*, *in* and *out*, *front* and *back*, *on* and *off*, *deep* and *shallow*, and *central* and *peripheral*. Thus, by gaining an understanding of the spatial associations of particles, L2 learners should be able to acquire PVs more easily than simply memorising their meaning (Haugh & Takeuchi, 2023:658; Yasuda, 2010:251). This viewpoint, therefore, considers the particle to be the provider of "rich imagery and schematic content", in contrast to the verb proper, which provides very little (Yasuda, 2010:252).

On the other hand, Mahpeykar and Tyler (2015:2) posit that both the verb and the particle in the *verb + particle* PV construction are polysemous in their own right, and that by studying "the interaction between the polysemy networks of the verb and the particle", it might be possible to derive stable,

less unpredictable meanings for PVs. PV combinations can, therefore, be thought of as "compositional" in nature, rather than random (Mahpeykar & Tyler, 2015:32). So, by investigating PVs such as *get out*, *get up*, *take up* and *take out*, they have been able to illustrate that, whereas the particle has previously been understood to be the polysemous agent in the *verb + particle* combination, the verb, in fact, likewise contributes a range of meanings. By considering the legitimate meaning options that the verb could have in combination with a particle which has its own meaning options, they illustrate that a PV construction "can have several systematic, motivated meanings" (Mahpeykar & Tyler, 2015:32). Whether either of these Cognitive Linguistic viewpoints will result in an easing of the multiple-meaning PV burden for L2 learners is yet to be seen.

2.7.5 Prolificacy

As early as 1920, PV formation was acknowledged as "a changing, growing tendency in language which throws up over night [sic], as it were, new combinations, and new meanings, so that an absolute and complete list would be impossible of realization" (Kennedy, 1920:5). Furthermore, Kennedy (1920:40) comments that even those deeply committed to maintaining a high standard of English "cannot fail to be impressed by the strength of this growing tendency in English". Bolinger (1971:xi) calls the PV "an outpouring of lexical creativeness that surpasses anything else in our language". A look at recent PV additions to the language, already familiar and in everyday use, illustrates this prolificacy: *chill out, freak out, log off/on, max out, scroll up/down, sex up, space out* (Schmitt & Redwood, 2011:173).

It has been suggested that native speakers are able to create new PVs with "ease and facility" (White, 2012:420), as native speaker familiarity with the use of particles makes the creation and understanding of new PVs a matter of course, a facility that is possibly not available to L2 learners (Darwin & Gray, 1999:66). Darwin and Gray (1999:66), consequently, list "the highly productive nature of the phrasal verb in English" as a factor that increases the difficulty of PV use for L2 learners, although, as no empirical evidence is offered in support of this statement, this suggestion can only be surmised. It might very well be that advanced English L2 students are able to form new PVs with as

much ease as L1 students do, athough this must remain a supposition without further research into this issue.

In the context of prolificacy, it is interesting to note that some of the PV examples that Kennedy (1920) provides (such as *sport up*, *whack up* and *tack down*) are no longer in use. Therefore, the coining of a PV does not necessarily guarantee its survival past a decade or so of popular use. This is not unexpected, as many vocabulary items fall out of fashion over time, requiring a constant update of the language user's mental lexicon.

2.7.6 Obscurity

As was seen in the discussion on types of PVs, PV meanings can vary "on a cline of transparency" from relatively straightforward (literal PVs, such as *come in*, *go out*, and *pass through*) to obscure (idiomatic PVs, such as *give up*, *make up*, and *butter up*) (Siyanova & Schmitt, 2007:119-120). The issue of obscurity of PV meaning is underscored by Bolinger (1971:xii), who declares that "[t]he phrasal verb is a floodgate of metaphor". White (2012:421) states that metaphorical (or figurative) meaning may be contained in the verb proper, or in the particle, or in both elements of the PV, which further complicates matters. The result of this obscurity of meaning often leads to avoidance of PV use by L2 learners who "prefer semantically transparent combinations over opaque units, if they find they must use them" (McPartland-Fairman, 1989:1) .

A further difficulty that compounds the issue of obscurity for idiomatic and aspectual PVs is their polysemous nature, the solution to which, it has been suggested, is learning the meaning of each combination of *verb + particle* (Myers, 2018:4; Riguel, 2014:113; Siyanova & Schmitt, 2007:120). Keeping the vast number of PVs in mind, it would be more effective to provide lists of only the most commonly used PVs and their meanings, as has been suggested by Liu (2011) and Liu and Myers (2020). However, this is by no means a straightforward task, not least because of the influence that register has on the meanings of PVs (Liu, 2011:680; Liu & Myers, 2020:6; Siyanova & Schmitt,

2007:121) 407. As illustration, Liu and Myers (2020:407) offer the PV *make up*, which, in speech and fiction, may mean *fabricate* or *put on cosmetics*, while, in academic writing, it primarily means *compose*. In fact, Siyanova and Schmitt (2007:121) state that the problem for L2 learners is "not so much choosing the verb form which carries the correct meaning, but rather choosing the verb which has the appropriate register, and which conforms to the expectations of the speech community".

A further suggestion is to teach PVs according to their categories (Bronshteyn & Gustafson, 2015:94). Hence, literal PVs can be taken at face value. For example, the meaning of the PV *come in* may be deduced from *come* (as in *to approach*) and *in* (as in *into*). However, considering the multiple meanings suggested for both *come* and *in* (18 for *come*, and 12 for *in*) in the *Merriam-Webster Online Dictionary* (2024), deducing the meaning of literal PVs might not be as evident as might be presumed. Next, it is suggested that idiomatic PVs should be memorised (Bronshteyn & Gustafson, 2015:14). This seems to be a suitable suggestion, given that the meaning of an idiomatic PV such as *make out* cannot be deduced from its parts. On the other hand, this might prove to be an arduous task, given the variety of meaning senses of *make out* (taken from the *Merriam-Webster Online Dictionary* (2024): *to complete*, *to grasp the meaning of*, *to form an opinion or idea about*, *to represent as being*, *to pretend to be true*, *to represent or delineate in detail*, *to see and identify with difficulty or effort*, *get along*, *and to engage in sexual intercourse*. Another suggestion that has been proposed for teaching idiomatic PVs is to encourage learners to draw images of such PVs in the classroom, an example of which, reproduced from Bronshteyn and Gustafson (2015:97), is shown here:

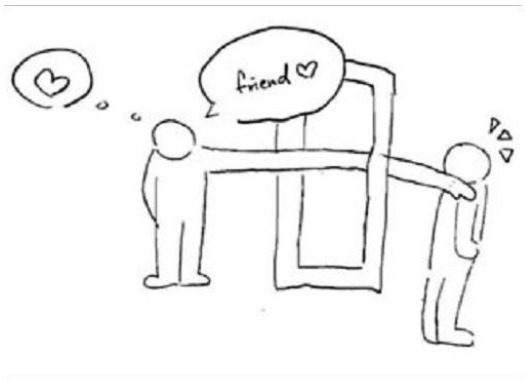

Figure 2.1 *Description of the PV reach out (Bronshteyn & Gustafson, 2015)*

Bronshteyn and Gustafson (2015:94) include the aspectual PV category in their suggestions for teaching PVs, which they suggest can be taught as having a literal verb proper but an idiomatic particle. However, the diverse opinions among researchers as to defining the aspectual PV (see §2.4) make proposals for its teaching problematic.

2.7.7 Influence of a mother tongue without a PV structure

The PV (along with other multi-word verbs) is a particular feature of the Germanic languages, to which English belongs (Celce-Murcia & Larsen-Freeman, 1983:265; Dagut & Laufer, 1985:78; Darwin & Gray, 1999:65; Morales, 2000a:11; Siyanova & Schmitt, 2007:120); Waibel (2007:160). Other languages that fall into this category include German, Dutch, and the Scandinavian languages (Celce-Murcia & Larsen-Freeman, 1983:265). Although the *verb + particle* structure is not entirely absent from non-Germanic languages, its use differs substantially (Liao & Fukuya, 2004:211). As a result, students whose mother tongue does not contain a multi-word verb structure might not develop strategies for using such structures (Alhatmi, 2023:4; Dagut & Laufer, 1985; Siyanova & Schmitt, 2007:120).

In a study that compared non-native to native PV use, levels of overuse and underuse were shown for various languages (Riguel, 2014:113-114). Two corpora were used, namely the International Corpus of Learner English (ICLE), comprising essays written by advanced L2 learners from a variety of countries, and the LOCNESS. The languages represented in the non-native corpus were Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish and Swedish, thus including languages that feature PVs (Germanic languages), as well as languages that do not. The resulting table (Table 2.2 below) is reproduced from Riguel (2014:114).

Table 2.2 Extent of over- and underuse of PVs by foreign learners in comparison to native speakers

Corpus	Percentage of use
LOCNESS (control corpus)	0%
ICLE German	+13.92%
ICLE Dutch	-0.1%
ICLE Polish	-0.87%
ICLE Finnish	-15.4%
ICLE Bulgarian	-17.59%
ICLE Swedish	-18.41%
ICLE Russian	-25.29%
ICLE French	-26.98%
ICLE Czech	-28.2%
ICLE Italian	-30.39%
ICLE Spanish	-44.57%

Referring to the information contained in the table, Riguel (2014:114) rightly points out that students whose mother tongue is German and Dutch (thus, two Germanic languages) do not show avoidance of PVs. Indeed, it seems that German students are more likely to use PVs than are native English speakers. This is in keeping with the idea that having a mother tongue containing a PV structure will result in fewer problems with PV use in English because L2 learners perceive their mother tongue to be similar to the language they are trying to learn (McPartland-Fairman, 1989:10). In line with the theory of mother-tongue influence, the results further indicate that non-Germanic languages, namely French, Italian, and Spanish (three Romance languages), show notable underuse (and, therefore, avoidance) of PVs. These results are confirmed by other studies: Waibel (2007) found the same discrepancy in PV use between German and Italian L2 learners; Zhou and Wang (2024:4) similarly observed overuse by German L2 learners, and underuse by Italian and French L2 learners. Furthermore, Zhou and Wang (2024:4) add Brazilian L2 students to the list of PV underuse, and Badem and Simsek (2021:64) add Turkish L2 students.

However, again referring to the table, it is interesting that speakers of Polish, not a Germanic, but a Slavic language, also did not show notable underuse of PVs. (The Slavic group of languages does not

have a PV construction.) On the other hand, speakers of Swedish, a Germanic language, showed substantial underuse of PVs (almost 20%). Zhou and Wang (2024:4) also mention that Malaysian L2 learners (Malaysian being a language that does not contain a PV structure) showed a similar pattern of high-frequency PV use to that of native speakers represented in the BNC, giving the reason for this dichotomy "frequent English exposure out of class". Therefore, it seems that one cannot state categorically that L2 learners whose mother tongue has a PV construction will have less of a problem with PV use in English than L2 learners whose mother tongue has no such construction, and vice versa.

The influence of the L1 on PV use has, indeed, been challenged. Chen (2013a:433), for example, observed no difference in PV use when comparing the writings of native (British) and non-native (Chinese) students, even when the PV construction is absent from the mother tongue of the L2 group. In fact, Chen (2013a:433) is of the opinion that competent L2 PV use observed in German students (as, for example, illustrated in Riguel (2014)) can be ascribed to the general English language competence of the L2 learners. Following on from this argument, she maintains that "the L1 factor is not the most influential parameter in phrasal verb acquisition; other factors may be of more importance" (Chen, 2013a:436). (The Chen (2013a) study is discussed in more detail in §2.9.7.)

The results of a study conducted by McPartland-Fairman (1989:100) also show no such inter-language influence. In addition, McPartland-Fairman (1989:3) questions the influence of a lack of a similar structure in the mother tongue on L2 learner problems by citing the complexity inherent in PVs. Moreover, there are several other factors that might negatively influence adult L2 learners who are attempting to learn a new language, such as the incorrect assimilation of new syntactic structures, the inability to incorporate new word meanings, and the influence of cultural norms that do not apply to the language being learned (McPartland-Fairman, 1989:3).

Nevertheless, some arguments about language similarity are worth noting. It has been suggested that language transfer takes place when L2 learners perceive their mother tongue to be similar to the language they are trying to learn (McPartland-Fairman, 1989:10). In spite of these similarities,

Bronshteyn and Gustafson (2015:93) warn that, even for L2 learners with Germanic mother tongues, language transfer will be problematic because of the idiomatic nature of some PVs and because "correspondence between languages" might lead to confusion rather than clarity. An example of this is the confusion between "false friends" commonly found in the vocabulary of different languages, such as the French word *sensible* and its English cognate which does not carry the additional meaning of *sensitive* that is conveyed by the French term.

Here it should be noted that Afrikaans, one of the 11 South African official languages, and for whom there will certainly be representation in the corpora, is a Germanic language. It may, therefore, be expected that Afrikaans students will not find the PV construction problematic. On the other hand, as will be seen in §2.9, and, as discussed in this section, having a similar construction in the mother tongue does not necessarily ensure competent use of the PV.

The influence on PV use of a mother tongue from which a PV structure is absent is discussed in more detail in §2.8.

2.7.8 Non-standard use

Apart from avoiding PVs, L2 students have also been found to make mistakes in the use of PVs, or to employ what is considered non-standard use. Non-standard uses could include, for example, the incorrect particle insertion in a PV (e.g. using *cut up* instead of *cut off* when meaning *interrupt*), or its entire absence (e.g. using *cut* when meaning *cut up*) (McPartland-Fairman, 1989:4). Mazaherylaghab (2015:93) also points out instances of hypercorrection, as in *return back*.

Furthermore, Zarifi and Mukundan (2014:52) speak of "unnaturalness", which is the result of "any learner production which is inappropriate", meaning that, even though language rules might have been applied correctly when producing an English phrase, the way in which the phrase is subsequently used is inappropriate. The result is language that sounds unnatural to the ear of the native speaker. To illustrate this point, Zarifi and Mukundan (2014:52) provide examples such as *'***break out** *the world*

record' and *'***come up** *the idea*' that illustrate instances where L2 learners have combined legitimate PVs with inappropriate noun phrases. A second type of unnaturalness is when an incorrect particle has been attached to the verb proper, as in *'*environmental changes* ***brought along*** *by mankind*', where *brought about* would have been the correct combination (Zhou & Wang, 2024:6). Zarifi and Mukundan (2014:60) include the overuse of PVs by L2 students, as well as the overrepresentation of normally infrequent PVs, as illustrations of unnaturalness.

Zhou and Wang (2024:5) also mention "creative" learner production of PVs, which refers to the use of PVs which are not found in dictionaries. Zarifi and Mukundan (2014:60) give *spend up* and *rescue up* as examples, where L2 learners have attempted to follow the pattern found in *eat up* and *drink up* to indicate completion. These examples demonstrate an attempt by L2 learners to use PVs in a productive manner, but without as yet sufficient familiarity with the L2 language. From this discussion, it would seem that there is a fine line between creativity and unnaturalness (Zhou & Wang, 2024:6). Waibel (2007) refrains from labelling instances of unnaturalness and creativity as errors, as, according to her, they indicate an understanding of the flexibility of PVs. (See §3.8.2.6 for a further discussion of how non-standard PV use will be reported on in this research.)

After consideration of all the concerns mentioned above, it is no surprise to find that the L2 learner is tempted to use coping mechanisms to overcome the problems of PV acquisition. Avoidance appears to be such a mechanism. In the next section, the use of avoidance to overcome the challenges that the PV presents will be examined.

2.8 Avoidance as an L2 strategy to overcome challenges with PV use

At this point, it is clear that the PV is both of substantial importance in the attainment of English proficiency, and significantly problematic in its acquisition (Garnier & Schmitt, 2015:658). It is to be expected, then, that under such circumstances, L2 learners will attempt to find an alternative that will not expose their linguistic inadequacies (Hulstijn & Marchena, 1989:242; Laufer & Eliasson, 1993:36). The reported prevalence of avoidance behaviour by L2 learners suggests that this is the coping

strategy of choice (Chen, 2013a; Dagut & Laufer, 1985; Hulstijn & Marchena, 1989; Laufer & Eliasson, 1993; Liao & Fukuya, 2004; McPartland-Fairman, 1989; Siyanova & Schmitt, 2007). Avoidance behaviour can be described as the avoidance of certain grammatical structures by non-native speakers of a language in order to overcome what is perceived by such learners as a challenge. For example, L2 learner preference for "semantically transparent combinations over opaque units, if they find they must use them" (McPartland-Fairman, 1989:1), can be seen as a sign of avoidance behaviour.

Researchers point out that avoidance does not necessarily equate to a lack of knowledge of the grammatical structure to be avoided, but rather to an awareness, though it might be slight, of such a structure (Hulstijn & Marchena, 1989:243; Laufer, 2000:186; Laufer & Eliasson, 1993:36; Siyanova & Schmitt, 2007:133). Indeed, awareness of the PV might be at the very root of the problem, as the multiple rules suggested for the identification of PVs are cumbersome and could be regarded as confusing to students (Darwin & Gray, 1999:66). Since the 1970s, as a result of important studies by Schachter (1974) as well as Kleinmann (1977), researchers have started to focus more on the avoidance of certain grammatical structures by L2 learners rather than solely on the errors made by these learners (Hulstijn & Marchena, 1989:242). In fact, it should be noted that avoidance does not necessarily lead to incorrect grammar usage, but rather to underuse of a structure that would be found to a greater degree in the language of native speakers, and overuse of more formal alternatives that might appear unnatural in some settings (Celce-Murcia & Larsen-Freeman, 1983:265; Darwin & Gray, 1999:65; Laufer & Eliasson, 1993:36). Darwin and Gray (1999:65), furthermore, point out that avoidance of PVs prevents students from becoming familiar with their use.

Dagut and Laufer (1985:73) indicate the importance of properly identifying "genuine avoidance phenomena" in order to "throw light on what would otherwise remain hidden recesses of uncertainty in the learner's mind". In addition, correctly identifying avoidance behaviour can contribute to understanding the processes that underlie L2 learning, as well as assisting in the creation of appropriate learning materials (Laufer & Eliasson, 1993:36).

Research into the causes of avoidance has produced two opposing views, namely intralinguistic and crosslinguistic. The intralinguistic view considers interference from the L1 as minimal, and presumes that difficulties within the L2, such as the "inherent complexity of the avoided item", as the principle cause of problems for the learner (Laufer & Eliasson, 1993:36-37). The crosslinguistic view encompasses more complexity. On the one hand, the absence of a similar structure in the L1 might be thought to result in avoidance behaviour, whereas, on the other hand, it might well be the existence of a similar structure in the L1 that confuses L2 learners and causes avoidance (Laufer, 2000:186; Laufer & Eliasson, 1993:37). Essentially, the crosslinguistic view posits that avoidance of PV use can be seen as L1 interference in L2 learning (Dagut & Laufer, 1985:73; You, 1999:149). Siyanova and Schmitt (2007:132) provide an illustration of this factor by citing a Spanish research participant, who, after being resident in both the UK and USA for more than eight years, "still felt more comfortable using one-word verbs 'of Latin origin'". On the other hand, Hulstijn and Marchena (1989:242) suggest that avoidance behaviour may be "doubly determined", meaning that factors influencing such behaviour should be sought on both sides of the L1 / L2 debate.

Laufer (2000:187) offers an extension to the crosslinguistic discussion. According to Laufer (2000:187), differences between languages should not be viewed simplistically as a "unitary phenomenon" based on a straightforward view of similarity and difference, but should rather be measured along three dimensions. The first dimension is conceptual: conceptual difference indicates the absence of a grammatical category in one of the two languages (such as the absence of article use in Russian), whereas conceptual similarity indicates the presence of a grammatical category in both languages (such as the grammatical gender subdivisions of male and female existing in Hebrew as well as French). The second dimension is formal: formal similarity indicates the expression of meaning in the two languages by means of the same grammatical structures (for example, in both English and French the passive voice is expressed using a similar structure), whereas formal difference indicates the expression of meaning by means of different grammatical structures (for example, in Hebrew, the passive voice is expressed using very different structures to those of English and French). The third

dimension is distributional: distributional difference indicates the use of a similar grammatical structure for different purposes or conditions (an example would be the difference of application of the passive voice in English and Hebrew), whereas distributional similarity indicates the use of a similar grammatical structure for the same purpose (for example, the relative clause describing the preceding noun in both English and Hebrew). (The examples used to illustrate the three dimensions are those of Laufer (2000:187)). Laufer's (2000) subsequent research made use of these three dimensions to compare L2 student use of four types of English and Hebrew idioms, representing various degrees of similarity. The results of Laufer's (2000) research (in this case, "partial formal similarity and distributional difference") seemed to support the hypothesis that avoidance depended on the varying *degrees* of similarity between the L1 and L2, since "the umbrella term L1-L2 difference covers more than one type of relationship between languages" (Laufer, 2000:187, 195).

It has also been suggested that English proficiency might have an influence on avoidance behaviour. For example, Hulstijn and Marchena (1989:250) claim that

> with moderate proficiency, learners feel tempted to adopt a play-it-safe strategy by using multi-purpose one-word verbs with general meanings, rather than restricted-purpose phrasal verbs with specific, sometimes idiomatic, meanings.

Research conducted by Schmitt and Redwood (2011:188) in England (with students both of English as a second language and English as a foreign language) and Wei (2021:35) (with Chinese L2 students), likewise, clearly show a relationship between proficiency levels and PV knowledge, as does research carried out by Haider *et al.* (2020:1194) (with Jordanian students).

A study by Riguel (2014:120) into an English-speaking child's acquisition of the PV, while not directly related to the present study, does, however, confirm the influence of proficiency. Her longitudinal study (following a child's language development from 11 months to three years, 10 months) showed that PVs were acquired in three stages, "from incomplete forms to complete constructions" (Riguel, 2014:120). Furthermore, the PVs most frequently used by the parent were picked up first by the child, and the PV types and frequencies used by the parent were emulated by the child (Riguel, 2014:121).

You (1999:150), reporting on PV avoidance by Korean L2 learners, adds another dimension to the issue of proficiency by claiming that "language learners have to reach a certain age to use semantically difficult structures (e.g. phrasal verbs) in their L2". This claim is based on the unexpectedly lower PV proficiency shown by Korean children who had been raised in the USA, compared to their parents who had been educated in Korea. You (1999:140) also finds three years of immersion in the L2 to be the threshold for adult PV proficiency. Further issues that might possibly influence PV proficiency could be educational methodology (in that, in Korea, PVs are simply taught as synonyms for their one-word alternatives, without any distinction in register or use), lack of context (in that the use of PVs in a natural setting is not illustrated), and the influence of the major field of study (in that Korean students who have English as a major are more likely to be invested in learning such a complicated grammatical structure) (You, 1999:151-152).

Lastly, it is also necessary to report that PV avoidance by L2 students is not the norm in all cases. For instance, Chen (2013a:425) reported a similar frequency of PV use between Chinese L2 learners and British native novice students. In some cases, L2 PV use was even found to be greater than native PV use. For instance, Singaporean L2 learners were observed to use more PVs than native speakers did, which was ascribed, in part, to the use of an informal writing style (Zarifi & Mukundan, 2014:53).

The most relevant findings of research into avoidance will be discussed in the next section.

2.9 Findings of previous research into avoidance

Previous research has investigated PV use from a variety of angles, ranging from the general functionality of PVs, such as the identification of the most frequently used PVs, to the main concerns with PV use, such as PV avoidance in L2 writing. It should again be noted that, as no South African research was available regarding phrasal verbs, international sources were referenced.

2.9.1 Dagut and Laufer (1985)

One of the first investigations into avoidance of PV use by L2 learners was conducted by Dagut and Laufer (1985), working with various groups of first-year university students at a Hebrew university, all of whom had had several years of exposure to English at school. The researchers approached the research from two angles: firstly, to determine whether the impression that such avoidance was taking place was, in fact, valid; and, if so, to establish the extent to which it was taking place. Using a multiple-choice test, they first identified the PVs that were preferred to one-word equivalents by the majority of the participating native speakers, resulting in a list of 15 PVs that included literal, figurative and aspectual types. These PVs were then used in multiple choice, verb translation and verb memorisation tests (using the same 15 sentences throughout, but adapted to the type of test) to determine the extent to which the L2 learners' choices aligned with those of the native speakers.

An example of the multiple-choice question was: "*We didn't believe that John could ever _____ his friends. (let down, solve, disappoint, carry on*". The four words at the end of the sentence included the correct PV (*let down*), its one-word alternative (*disappoint*), and two distractors (*solve* and *carry on*). Students were asked to complete the sentence using the verb most likely to be used in spoken English (therefore, the most colloquial). For the verb translation test, an example sentence read: "*We didn't believe that John could ever _____ his friends (leachzev)"*[3]. The infinitive form of the Hebrew verb that needed to be translated to complete the sentence was supplied at the end of the sentence. Finally, the sentence: "*Johnny argued and refused to give in to his older brother (Johnny hitvalkeach veserav lehikana leachiv hagadot)*" was an example of the verb memorising test. The sentence, along with four distractor sentences, each containing one-word alternative verbs, were shown to the students. An hour later, they were asked to complete the original sentence, from which the verb had been removed. (The examples are all taken from Dagut and Laufer (1985:74-76)).

[3] Hebrew transcriptions are provided in Roman script.

The results of the tests indicated that the L2 students preferred the use of one-word equivalents over PVs, and that, where PVs were used, the preference was for literal PVs. For example, in the multiple-choice test, completed by 60 EFL students, 72% of the literal PVs were selected, 48% of the aspectual PVs, and only 27% of the figurative PVs, resulting in an average of 58% of PVs selected. For the translation test, a different group of 60 students was used, half of whom were students of English language and literature and were, therefore, assumed to be proficient in English. The English majors used PVs in 32% of their translations, of which 53% were literal, 33% aspectual, and 21% figurative. The remaining 30 (EFL) students doing the translation test chose PVs in 15% of their translations, of which 16% were literal, 22% were aspectual, and 11% were figurative. Fewer PVs were thus used overall in this test, despite the fact that half of the participants were proficient in English. On the other hand, as Dagut and Laufer (1985:76) point out, students required more knowledge of English to complete this test than for the multiple-choice test. A third group of 60 EFL students completed the verb memorisation test. Completed sentences using PVs were provided for students to study, on which they were tested an hour later, the PVs having been removed. Even though students had been made aware of the correct PVs within the context of these sentences, PVs were only selected in 24% of the responses, of which 37% were literal, 35% aspectual, and 13% figurative. As there was evidence of students having knowledge of PVs, both because these are taught in the English curriculum in Israel, and since PVs were occasionally selected by students, these results indicate "genuine avoidance rather than … ignorance" (Dagut & Laufer, 1985:77-78). For Dagut and Laufer (1985:78), such avoidance is the result of Hebrew having no equivalent grammatical structure and is, therefore, "corroboration of the dominant role of L1 in the L2 learning process".

According to Liao and Fukuya (2004:198), relying on intuition rather than research, as Dagut and Laufer (1985:75) did, by assuming that the students had come across the 15 PVs during their schooling, might have resulted in the incorrect interpretation of results. In addition, Liao and Fukuya (2004:198) point out that, in drawing their conclusions, Dagut and Laufer (1985:78) concentrated on the L1-L2 (crosslingual) influence only, without also referring to the difference between literal and figurative PV

avoidance. For Liao and Fukuya (2004:198), the evident avoidance of figurative PVs rather than literal PVs indicates that an intralingual influence, and not only a crosslingual influence, is at play when avoidance takes place.

2.9.2 Hulstijn and Marchena (1989)

In a follow-up study to Dagut and Laufer (1985), Hulstijn and Marchena (1989) conducted research aimed at clarifying whether the avoidance identified in the previous study resulted from semantic causes (the figurative nature of PVs) or syntactic causes (interference from the L1). This research was, therefore, designed to challenge the supposition underlying Dagut and Laufer's (1985) study that PVs would not be avoided if the structure existed in the L1. Additionally, Hulstijn and Marchena (1989) also surmised that avoidance would decrease as proficiency increased. As they were interrogating the findings of the Dagut and Laufer (1985) research, the design of the study was similar, but not identical. The researchers acknowledged the influence that task types have on test results in the varying degrees to which PVs and one-word alternatives are made available to the learners taking the tests. Consequently, data were collected using a variety of tasks, following the example of Dagut and Laufer (1985). However, the study differed in various ways, such as dividing learners more precisely into intermediate and advanced groups, and using different sentences from those in the original study. Fifteen pairs of PVs preferred by native English speakers, with matching one-word alternatives, were identified for use in the tests. Learner familiarity with these verbs was confirmed by teachers and textbooks. Furthermore, participants were asked to indicate any verbs in the tests that were unfamiliar to them, to which there was no response.

The intermediate participants, 17-year-old students with five to six years of schooling in English, made up three of the test groups, and the advanced learners, first-year university or teacher-training students who were studying English, made up another three groups. The multiple-choice task (in which both PVs and their one-word equivalents were made available) was completed by 50 of the intermediate and 25 of the advanced L2 learners, the memorisation task (in which only PVs were made

available) by 50 learners from both groups, and the translation task (in which neither PVs nor one-word equivalents were made available) by 25 learners from both groups. The multiple-choice and translation tasks included the instruction to "[a]ssume that these sentences have been written in normal, colloquial English". This instruction was presumably meant to encourage the use of PVs over one-word alternative verbs.

If avoidance behaviour was to be found, it was expected to surface most clearly in the memorisation test, where the appropriate PVs had been provided. This proved not to be the case. Both groups of learners generally recalled the PVs originally provided, and did not use one-word verbs in preference to PVs. The intermediate group did, however, make more mistakes than the advanced group. The results of the multiple-choice test indicated that the intermediate learners used far fewer PVs than the native speakers, while the advanced learners had a pattern of PV use very similar to that of the native speakers. Finally, the results of the translation test showed that the advanced students had used a similar number of PVs and one-word verbs, which is an unexpected result, given that the PVs were not made available in this test as they were in the multiple-choice test. The intermediate learners showed a preference for one-word verbs, although, at the same time, they used a substantial number of PVs. As is to be expected, the advanced learners showed a higher level of proficiency in PV use than the intermediate learners. Nevertheless, both groups displayed avoidance behaviour as far as four particular PVs (*give up*, *break out*, *go off*, and *bring up*) were concerned, possibly because these PVs were very similar to their Dutch equivalents.

According to Hulstijn and Marchena (1989:250), avoidance behaviour was manifested, to a greater or lesser degree, on three occasions, namely when the L2 grammatical structures differed greatly from the L1 grammatical structures, when equivalence between the L1 and L2 might have created suspicion about the validity of direct transfer, and when the PVs were figurative in nature. The first two observations indicate crosslinguistic influences, and the third an intralingual influence. While these observations were made for both the advanced and intermediate students, the advanced students

showed greater ability in the use of PVs on the whole, indicating the influence of language proficiency. As the Dutch language also has a PV construction, Hulstijn and Marchena (1989) argue that avoidance cannot be linked to lack of familiarity with the construction. Consequently, in response to the findings of Dagut and Laufer (1985) discussed previously, Hulstijn and Marchena (1989:251) suggest that avoidance behaviour might be linked to semantic in addition to purely syntactic reasons.

2.9.3 Laufer and Eliasson (1993)

Laufer and Eliasson (1993) conducted research based on that of Dagut and Laufer (1985), and in response to that of Hulstijn and Marchena (1989). They were particularly interested in finding out how the traditionally held causes of avoidance (differences between the L1 and L2, similarities between the L1 and L2, and the inherent complexity of the L2) intersect, if at all (Laufer & Eliasson, 1993:37). While they were interested in avoidance of L2 grammatical structures in general, they based their research predominantly on PV use. Two tests (a multiple-choice and translation test) were written by 87 advanced adult Swedish-speaking L2 learners (with 50 of the students writing the multiple-choice, and 37 the translation test). Swedish, like English, contains PVs. Laufer and Eliasson (1993) started out with the same presumption of PV familiarity as Dagut and Laufer (1985), based on their knowledge of the inclusion of high-frequency PVs in the school curriculum. However, they included an independent test to confirm that presumption, thus avoiding the potential interpretation problem mentioned by Liao and Fukuya (2004:198).

The participants were instructed to use colloquial English. The aims of the study were, firstly, to see if there is a pattern of avoidance of English PVs in general; secondly, to see if avoidance behaviour was more evident where the English PVs were different from the Swedish PVs as opposed to English PVs for which there are Swedish equivalents; and, thirdly, whether there was a more marked avoidance of figurative PVs than of literal PVs in English. The investigation included comparison with the results generated by the Hebrew students in the Dagut and Laufer (1985) study. This inclusion strengthened the validity of the study for two reasons: the two groups were comparable as far as level of proficiency

in English was concerned, but differed as far as the presence of PV construction in the mother tongue was concerned.

From the results of the tests, the following conclusions were drawn by the researchers: PVs are not avoided by L2 learners if such structures exist in the mother tongue, in contrast to L2 learners whose mother tongue does not contain the grammatical structure. Furthermore, figurative PVs are not avoided by Swedish-speaking L2 learners because of "idiomatic disbelief". This refers to the phenomenon also reported by Hulstijn and Marchena (1989:250), that "similarity in form combined with similarity in (idiomatic) meaning between an L1 and an L2 item may lead learners to avoid using the L2 item".

Lastly, the complexity of the L2 does not seem to be a significant factor in the avoidance of PVs, although "[t]hat is not to say that it has no effect at all. If a structure is complex and missing in the L1, it is a better candidate for avoidance than a simple structure that is absent in the L1" (Dagut & Laufer, 1985:44). In contrast to the findings reported by Hulstijn and Marchena (1989), the final conclusion drawn by Laufer and Eliasson (1993:46) is that grammatical differences between the mother tongue and the target language have the greatest effect on avoidance behaviour. Although the researchers added a proviso that results might differ for L2 learners at lower levels of proficiency, it is clear that they believed the influence of "L1-L2 difference" to be true regardless of proficiency level.

2.9.4 Liao and Fukuya (2004)

Liao and Fukuya (2004) investigated possible PV avoidance behaviour in Chinese L2 learners. While the Chinese language also has a *verb + particle* structure, it functions differently from the PV structure found in English. The participants in the study consisted of a group of 15 English mother-tongue speakers, and 70 advanced and intermediate Chinese L2 learners, divided into six groups. By including both advanced and intermediate L2 learners, the researchers hoped to address the influence of English proficiency on avoidance. While their test materials were based on those of Dagut and Laufer (1985), different PVs were used, for the following reasons. Firstly, the researchers were concerned

that the PVs used in the previous studies might not be appropriate in an American setting. Secondly, as PVs are often used colloquially, the sentences used in the tests (multiple-choice, verb translation, and recall) were shortened and given a more colloquial flavour than those of the Dagut and Laufer (1985) and Hulstijn and Marchena (1989) studies, thus ensuring an informal environment suitable to PV use. Using mother-tongue speaker preferences (established by means of a multiple-choice test written by the group of native speakers), 15 PV and one-word equivalent pairs were selected for use in the test. It should be noted that 11 of the 15 PVs used in this study were figurative, which, as the researchers later conceded, could have affected avoidance behaviour. This is because of the acknowledged semantic difficulty presented by figurative PVs (Dagut & Laufer, 1985; Laufer & Eliasson, 1993), namely that "the meaning of the verb departs from the meaning of its individual components" (Liao and Fukuya (2004:215). Each of the six groups of Chinese L2 learners completed one of the three tests, namely multiple-choice, translation or recall.

The findings of the study showed, firstly, that, of the two groups of Chinese L2 learners, only the intermediate students were inclined to avoid PV use, preferring to use one-word alternatives. The advanced L2 students showed a very similar pattern of PV use to that of the native English speakers, with only a slightly lower use of PVs. Secondly, as "the different frequency of figurative and literal phrasal-verb usage found for the non-native speakers is consistent with the frequency of phrasal-verb production by the native speakers" (Liao & Fukuya, 2004:215), it can be said that all three groups preferred to use literal rather than figurative PVs. However, actual avoidance behaviour was only seen in the intermediate L2 learners since the advanced learners were shown to have the same pattern of literal and figurative PV use as the native speakers. Apart from the acknowledged semantic difficulty of figurative PVs, Liao and Fukuya (2004:215) concede that the distribution of PV types, 11 of the 15 PVs being figurative and only four literal, might have affected results. A third research question investigated the effect of test type on results, which showed that it was the translation test in particular that resulted in a preference for literal PV use over figurative PV use. As neither a PV nor a one-word alternative is made available in this test, the researchers surmise that "inherent L2

complexity" was at play (Liao & Fukuya, 2004:216). Their research, then, points to three factors that affect PV avoidance, namely proficiency, PV type and test type.

According to the researchers, thus, it is reasonable to link avoidance to the lack of a similar structure in the L1 language, which is overcome as familiarity with the L2 increases (Liao & Fukuya, 2004:211), a supposition underscored by the Hulstijn and Marchena (1989) study. In fact, Liao and Fukuya (2004:213) contend that previous research (Dagut & Laufer, 1985; Laufer & Eliasson, 1993) would have pointed to the same conclusion had proficiency levels been incorporated into the studies: even if PV avoidance behaviour is evident, especially where the L1 does not have a similar lexical structure, such avoidance dissipates as English proficiency increases. Liao and Fukuya (2004:214) further point out that L2 learners who are in an immersive environment (for example, are studying in North America) might have greater familiarity with PV use because of its supposed colloquial nature. This point is strongly indicated by the fact that, in their own study, most of the intermediate learners had had limited exposure to English, whereas the advanced students had spent between nine months and three years "in a native English environment" (Liao & Fukuya, 2004:213). The researchers deduce that the prevalence of PVs in informal settings would make such an environment ideal for acquiring PV use.

2.9.5 Siyanova and Schmitt (2007)

The research conducted by Siyanova and Schmitt (2007) did not focus specifically on PV use, but rather on that of all multi-word verbs. Nevertheless, in incorporating PVs in their research, their findings are relevant here, as they investigated various aspects of multi-word verb use, such as the comparison of spoken and written discourse, and native and non-native use (Siyanova & Schmitt, 2007:121). Comparing non-native to native multi-word verb use is especially significant, as avoidance of a specific structure can only be identified "if there is evidence that native speakers would use the form that we claim is being avoided by L2 learners in the context under consideration" (Siyanova & Schmitt, 2007:133). They also looked at whether extended exposure to environments where multi-word verbs

are extensively used would improve the multi-word verb use by L2 students, thus exploring the effects of language proficiency.

In contrast to the studies previously discussed, data from corpora rather than tests were used, although data from questionnaires were also included. Furthermore, the researchers incorporated both speech and written texts in their research. A further difference is the method used to select the multi-word verbs for the research: 14 multi-word verbs were selected from the research conducted by Biber *et al.* (1999), Dagut and Laufer (1985) and Liao and Fukuya (2004), which suggests that they were mostly, if not all, PVs. A further 12 were added on a more arbitrary basis[4]. These were a variety of types of multi-word verbs that they had encountered while planning their research. The most appropriate one-word alternative verbs were then selected, although, as Siyanova and Schmitt (2007:122) admit, such alternative verbs are sometimes, at best, "roughly synonymous". This supports the point made earlier by Thim (2012) about PVs not necessarily having feasible synonyms (§2.5).

The research consisted of two parts, the first involving corpora and the second a questionnaire. The corpora used were the Cambridge and Nottingham Corpus of Discourse in English (CANCODE) for the spoken discourse data of native speakers, the BNC for the written texts of native speakers, and the International Corpus of Learner English (ICLE) for the written texts of advanced L2 learners. The frequency of use of the 26 multi-word verbs were searched for in these corpora. The "questionnaire" that was included in the research was, in fact, a type of multiple-choice test, presented in a colloquial style that encouraged PV use. An original version of the questionnaire, containing all 26 multi-word and one-word verbs, was completed by the native speakers, which resulted in a final version comprising the verbs that had been preferred by the native speakers. The questionnaire also included a Likert scale rather than a straightforward choice of responses, to allow learners to indicate a

[4] The researchers state that the further twelve multi-word verbs were "arbitrarily taken from a variety of texts and conversations", although, further along, they say that "the 26 verb pairs in the study were not randomly chosen", which appears somewhat contradictory (Siyanova & Schmitt, 2007:122, 125).

preference for more than one of the available choices, a factor that the researchers felt had been a shortcoming in previous tests. The researchers presumed participant knowledge of the selected multi-word verbs, although there was no formal test to confirm this (Siyanova & Schmitt, 2007:133).

Those completing the questionnaire were made up of 66 native and 65 non-native speakers (after unusable questionnaire responses had been discarded). The group of 66 native speakers consisted of undergraduate and postgraduate students, as well as graduated students who were working professionally. The non-native speakers were within the same range: 40 were undergraduate and postgraduate students, and 25 were graduated students who were working professionally, either in the UK or in their home countries. These non-native speakers were selected from those whose native tongues were non-Germanic, and, who, therefore, did not have a multi-word verb structure (Siyanova & Schmitt, 2007:122). It should be noted that, while the L1 languages of the non-native speakers who completed the questionnaire were non-Germanic, this is not necessarily true for the ICLE data, where 11 L1 languages were represented. This factor could have influenced the results that follow.

The results for the native speakers were unexpected in that the data from the two corpora indicated that, although multi-word verbs were used, one-word alternatives were preferred in both speech and written discourse. Significantly, in their analysis, the researchers include the proviso: "[t]o the extent that [the 26 selected multi-word verbs] are representative" (Siyanova & Schmitt, 2007:131), which allows for these results to be questioned. In fact, the results were somewhat contradicted by the data from the questionnaires that were completed by the 66 native speakers, which suggested that they preferred to use multi-word verbs in spoken discourse.

The results from the ICLE showed the same inclination by non-native speakers to use one-word alternatives in written discourse as the native speakers, although, on the other hand, the non-native speakers also used the one-word alternative verbs more frequently than did the native speakers. The researchers further compared the native speakers' use of multi-word verbs in speech (using the CANCODE data) to the non-native use of multi-word verbs in written discourse (using the ICLE), and

found the native use of multi-word verbs far higher than that of the non-native speakers, with mixed results for one-word alternative use. As the researchers themselves acknowledge, these results were not necessarily noteworthy because they were not comparing "like-with-like" (Siyanova & Schmitt, 2007:127). The preference for one-word alternatives by non-native speakers was confirmed by the data garnered from the questionnaires (Siyanova & Schmitt, 2007:131). The researchers were surprised to find that advanced non-native speakers seemed to avoid using multi-word verbs even in a colloquial setting, as this contradicted previous research. Hence, they compared exposure to an English environment among the 65 non-native speakers, and found only a slight inclination by those whose exposure was 12 months or longer for a multi-word verb preference. Siyanova and Schmitt (2007:131) conclude that the difficulty that multi-word verbs pose for learners of English is underscored by these outcomes.

The following three studies had different, though still pertinent, foci to those discussed above, and are, accordingly, presented separately. The McPartland-Fairman study (1989) examined the way in which PV meaning is accessed by L1 and L2 students, while the two Chen studies (2013a, 2013b) compared the PV use of Chinese students to that of British and American students.

2.9.6 McPartland-Fairman (1989)

In her study, McPartland-Fairman (1989) investigated how the meaning of PVs (whether literal, completive or figurative) is accessed, by comparing how native and non-native speakers approach PVs. The 32 non-native speakers (along with 32 native speakers) that took part in the test were all advanced English learners who could be expected to have a good understanding of PVs. Three pre-tests were used to establish the PVs that were to be used in the main test. To check that they were familiar with both the literal and figurative meanings of a PV, non-native speakers were asked to paraphrase the PVs provided, and native speakers were asked to identify the most generally used PVs in the English language. The PVs were also rated for semantic transparency. In the main test, the native and non-native speakers were tested on how they accessed literal and figurative PVs. Each participant heard a

sentence containing a PV by means of auditory input, and, at the moment the PV appeared in the sentence, saw a word flash on a screen, which the participant had to say aloud. The words on the screen could be literal or figurative alternatives to the PV, or simply control words with no links to the PV. The time lag between the appearance of the word on the screen and its pronouncement was measured to see if there was a difference between how L1 and L2 students accessed PV meanings, as well as to see if they accessed literal and figurative PVs differently. The results showed that, though non-native speakers accessed literal and figurative PV meanings slightly more slowly than native speakers, they accessed PV meanings as thoroughly and automatically as native speakers did (McPartland-Fairman, 1989:103-105). Furthermore, the results seem to indicate that non-native speakers access figurative PVs as single, lexicalised units, instead of first grappling with the meanings of the two separate items. This is in keeping with the assertion by Bronshteyn and Gustafson (2015:93) that figurative PVs are stored as units (See §2.7.1). Thus, advanced English learners do not seem to have any difficulty with the initial access to PV meanings, even when they are figurative (idiomatic). However, one of the pre-tests had indicated difficulty in paraphrasing PVs (McPartland-Fairman, 1989:83-84), which suggests that the results from the main test do not show the full picture. McPartland-Fairman (1989:104-105), thus, surmises that the problem arises further down the line from the initial access, when context needs to be considered in one's choice of PV.

2.9.7 Chen (2013a)

Chen (2013a) compared the PV use of Chinese students to that of American and British students. Data for the Chinese students were taken from a three-year longitudinal Chinese student corpus, comprising 780 argumentative essays, which had been collected twice a year under exam conditions, resulting in six sets of data (Chen, 2013a:421-422). For comparison, the argumentative essays of American L1 university students in the Louvain Corpus of Native English Essays (LOCNESS-US) were used, as well as the argumentative essays of British school leavers in the General Studies corpus (GS-UK).

The two comparison corpora were suitably comparable in most regards, but not in all. The British students were slightly younger, and the American students slightly older, than the Chinese students. The topics on which the essays had been based were similar across all the corpora. The British student essays, as with the Chinese essays, had been collected under exam conditions, but this was true for only some of the American essays. Finally, essay length differed substantially, with an average length of 242 tokens for the Chinese texts, 846 tokens for the American texts, and 675 tokens for the British texts. Wordsmith Tools was used to extract all multi-word verb combinations in each corpus, from which, by means of a manual process, a list of legitimate PVs was extracted per corpus. These lists formed the basis for the comparison between the three corpora.

Chen (2013a:435) makes an interesting observation about the underuse of PVs in that her research shows a significant difference in PV use between American and British native speakers, the American speakers being far more prolific in their use of PVs than the British speakers. This suggests stylistic differences between two groups of native speakers. Such variation in PV use by native speaker groups can have a marked effect on research into over- and underuse. As a result, the Chinese L2 student participants in her study showed underuse of the PV when compared to the American L1 speakers, but neither over- nor underuse when compared to the British L1 speakers. A further important observation by Chen (2013a:433) is that "learners' avoidance behaviour observed under experimental conditions does not correlate with considerable underuse of phrasal verbs in actual writing", which suggests that L2 learners might be inclined to select one-word alternative verbs over PVs in artificial settings. The participants in her study also showed a generally high frequency of verb use, which could account for the relatively high frequency of PV use (Chen, 2013a:434). Finally, Chen (2013a:434) observes that the frequency with which L2 learners encounter PVs plays a role in PV acquisition and production. This finding is supported by Schmitt and Redwood (2011:184), who found "a general trend of higher frequency leading to a greater chance of learning phrasal verbs to a productive degree of mastery".

2.9.8 Chen (2013b)

In a longitudinal corpus-based study, Chen (2013a) investigated the progress of Chinese undergraduate student PV use over the three years of their degree. (The corpus was also used in the Chen (2013a) study discussed in §2.9.7, and was described in detail there.) All PV tokens were extracted for the three-year period, and then compared. This study revealed illuminating results as far as the effect of English proficiency on PV use is concerned. Chen (2013b:97) found that there was no marked increase in PV use over the three years, and that, in fact, PV use decreased over that time, with a "considerable drop" in the second year. Chen (2013b:97-98) offers three possible explanations for these unexpected results: firstly, that proficiency is not necessarily linked to PV knowledge, due to the problematic nature of PV use; secondly, that the over-emphasis of PVs during examination preparation had possibly resulted in avoidance; or, thirdly, that students had become aware of the PV reputation of informality and had, consequently, started to avoid them. Of the three options, Chen (2013b:98) finds this third option unlikely, because of "learners' poor stylistic knowledge", as evidenced in the same study that shows low use by the Chinese L2 learners of high-frequency PVs. The results of this study suggest that conclusions drawn elsewhere regarding English proficiency and PV use should be regarded with caution. As Chen (2013b:98) observes, "PV acquisition is a complex process. It may involve considerable progress at certain phases as well as great regression at others".

The various studies on avoidance discussed above provide the following, sometimes contradictory, information. Firstly, some of the results seemed to confirm that avoidance has a crosslinguistic source, in that an L1 without a PV construction meant that the L2 learner would struggle to acquire PV use (Dagut & Laufer, 1985; Laufer & Eliasson, 1993; Liao & Fukuya, 2004). Secondly, in cases where avoidance behaviour was observable even where the L1 had a similar structure, some researchers equated a syntactic reason to the avoidance, namely the obscurity of figurative PVs (Hulstijn & Marchena, 1989). Thirdly, avoidance behaviour was, in fact, not always evident, or not evident to a notable degree. This was most often observed in advanced L2 students (Laufer & Eliasson, 1993; Liao

& Fukuya, 2004), although not in all cases, which was taken to confirm the complexity of acquiring PVs (Chen, 2013b; Siyanova & Schmitt, 2007). Furthermore, it appears that L1 PV use might differ across varieties of English, as, for example, in the case of British and American PV use (Chen, 2013a).

2.10 Conclusion

This literature review, firstly, provided some background as to the origins of the PV, thereby preparing the ground for some of the problems of PV use that emerged during further discussions. Thereafter, the issues surrounding the defining of the PV were illuminated, and the definition to be used in the current research established. Certain aspects pertinent to an understanding of the PV were then explained: the various PV categories which facilitate the use of PVs, the use of PVs across registers, and the importance of PV frequency in PV attainment. As the problematic nature of PV acquisition is the underlying factor that emerges from all these discussion points, the main concerns were then listed and briefly discussed. Finally, avoidance behaviour, the primary strategy employed by L2 learners in dealing with these issues, was introduced, and research focussing on this strategy listed and expounded, along with a few studies that examined the development of proficiency in PV use more broadly.

Chapter 3: Methodology

3.1 Introduction

In the methodology chapter, the procedures that will be used to address the research aims are explained. Any research must start with a clarification of the problem on which the research is based. A detailed outline of the methods by which the problem statement is to be addressed is given, and includes an explanation of the theoretical framework and research design in which the research is embedded. The discussion of the processing of the data includes a brief overview of the pilot study. Thereafter, the structure of the main study is explained, such as the corpora to be used, and the procedures that will be used to collect and analyse the data.

3.2 Problem statement

The overall purpose of this research study is to form a comprehensive view of PV use in English in the South African context. Necessarily, besides reporting on the observed patterns of PV use that emerge, PV use anomalies will also be indicated. Even though research indicates a link between correct PV use and achieving competence in English, the use of PVs by either South African L1 or L2 students has not yet been rigorously investigated, with the result that there is no clarity on possible issues regarding PV use by L2 students in particular. Should such issues be identified, they can serve as a means of adjusting and updating pedagogical material so that L2 learners are supported more effectively (Laufer & Eliasson, 1993:36). L2 student PV use cannot, however, be investigated in isolation, and will not lead to the comprehensive view that is aimed at here. For this reason, PV use by South African students who speak English as a first language (L1) will also be investigated and compared to L2 PV use.

3.3 Research aims

Flowing from the problem statement, the research aims are:

- To determine how L1 students use PVs in their writing.

- To determine how L2 students use PVs in their writing.
- To determine whether there is a difference in PV use between L1 and L2 students.
- To determine whether there is any difference in patterns of PV use across universities according to the ranking of the universities.
- To determine how patterns of PV use identified in the study could aid in the teaching of PV use to L2 students.

3.3.1 Research questions

Each research aim is addressed through the investigation of a related research question. The first set of these questions is designed to examine the pattern of PV use in L1 student writing:

1a. Which PVs are predominantly used by L1 students?

1b. To what extent do L1 students show a preference for PVs or for one-word alternatives?

1c. To what extent do L1 students adhere to syntactic and semantic norms in their use of PVs?

The second set of questions is designed to examine the pattern of PV use in L2 student writing:

2a. Which PVs are predominantly used by L2 students?

2b. To what extent do L2 students show a preference for PVs or for one-word alternatives?

2c. To what extent do L2 students adhere to syntactic and semantic norms in their use of PVs?

There are four further questions. The first of these is designed to examine the differences in PV use that are evident between L1 and L2 students. This is followed by a question that is designed to determine whether the PV use at a university can be aligned to its ranking. The final two questions are aimed at investigating whether the patterns of PV use that are identified in the research could be used to aid in the teaching of PV use to L2 students.

1. What are the main differences in the use of PVs by L1 and L2 students?

2. Does the occurrence of PV use differ across universities according to the ranking of the universities?

3. How could raising awareness of PVs and their alternative one-word verbs help students make appropriate choices in their academic writing?

4. How could reviewing errors in the use of PVs serve to guide students on how to use PVs correctly?

3.4 Corpus linguistic approach

The findings of the various research studies discussed in the literature review suggest that L2 attainment of competence in English is negatively influenced by unfamiliarity with and avoidance of the use of PVs. Consequently, it is assumed that similar patterns of PV avoidance will be observed in South African L2 writing. In order to assess the validity of this assumption, this research study follows a corpus-based approach. Biber *et al.* (1998:4) broadly describe the characteristics of a corpus-based approach as the use of electronic means to analyse the patterns found in large corpora of actual (natural) texts which have been ethically sourced. Rayson's (2008:519) description is more exact in that he states that the corpus-based approach starts from the identification of a research question that focuses on how a particular grammatical feature or construction is used. To put it more succinctly, a definitive distinction of the corpus-based approach is that existing theories are validated by means of corpus data (Botha, 2012:73-74). Indeed, a criticism of the corpus-based approach has been the accusation that "inconvenient evidence" is discarded to force the results to fit the assumption driving the research (McEnery, Xiao & Tono, 2006:8), although such a claim seems attributable to research ethics rather than methodology. Both quantitative and qualitative techniques are used in the analysis. According to Biber *et al.* (1998:4), the incorporation of qualitative interpretations of the patterns that emerge allows corpus-based analysis to do more than merely count linguistic features. Qualitative analysis is also necessary for an understanding of the significance of the findings.

The theoretical framework of this study is thus based on corpus linguistics, which is "a scientific method of language analysis" (Brezina, 2018:2). Biber and Reppen (2015:1) define corpus linguistics as "a research approach that facilitates empirical investigations of language variation and use, resulting in research findings that have much greater generalizability and validity than would otherwise be feasible". According to McEnery & Hardie (2012:1), corpus linguistics is not directly related to the study of language, but rather to the procedures used in doing so.

Because of the current availability of corpora "of suitable size and machines of sufficient power to exploit them", the possibility of investigating previously daunting language theories is greatly increased (McEnery & Hardie, 2012:1). Furthermore, the software tools that are available for corpus research vastly enhance analysis, such as concordances (used in qualitative analysis) and frequency list generators (used in quantitative research) (McEnery & Hardie, 2012:2). According to Biber and Reppen (2015:2), corpus linguistic research "offers strong support for the view that language variation is systematic and can be described using empirical, quantitative methods". This methodology requires that the researcher "provide empirical evidence in the form of data drawn from language corpora in support of any statement made about language" (Brezina, 2018:2). It is, therefore, a quantitative methodology, which means that descriptive statistics will necessarily be used to make sense of the collected data.

Quantitative and qualitative analysis will be used in this research. As expounded by Zhou and Wang (2024:4), quantitative analysis refers to, for example, investigating PV frequency and lexical diversity (PV types to tokens), and reporting the most frequently used PVs, as well as PV percentages "to showcase the common/uncommon presence of PVs". Qualitative analysis is used to "explore the quality of the PVs produced by L2 learners", and may, for instance, be used to identify unnatural PV use (Zhou & Wang, 2024:5).

3.5 Research design

The research design most appropriate for this research study is descriptive research as this kind of research is used "to describe the existing phenomena as accurately as possible" without intervention from the researcher, because "[t]he phenomena observed in descriptive research are already available" (Atmowardoyo, 2018:198).

According to Atmowardoyo (2018:198), descriptive research may include quantitative as well as qualitative data collection and analysis. As an illustration, he uses the example of data collected through questionnaires, which, while being qualitative in nature, still require quantification. Quantitative and qualitative research methods will be integrated in this research.

Quantitative research deals with numbered data that is analysed statistically. It is objective and deductive in nature, and it is important that its findings are generalisable and replicable (Creswell, 2009). Qualitative research, on the other hand, endeavours to explore and understand individual meaning, and analysis of the data depends on the researcher's interpretation (Creswell, 2009). It is, therefore, subjective and inductive in nature. Integrating these two methods, in what is known as mixed methods research, "involves the use of both approaches in tandem so that the overall strength of a study is greater than either quantitative or qualitative research" (Creswell, 2009:4). The section on data collection and processing (§3.8.2) will demonstrate how quantitative and qualitative data are integrated in this study.

It is necessary to clarify some of the terms that will be used. The use of the acronym *PV* for a phrasal verb has already been established. The one-word alternative verb that functions as an alternative to a PV will be indicated by the acronym *ALT*. The use of other acronyms, such as abbreviations for educational institutions (as in NWU for North-West University) are provided in the text where relevant.

The term *frequency* will be used when discussing the number of occurrences of a linguistic element. According to McEnery *et al.* (2006:52), in corpus linguistics, "frequency refers to the arithmetic count of the number of linguistic elements (i.e. tokens) within a corpus that belong to each classification scheme". According to Zhou and Wang (2024:4), the "most common direction in quantitative investigation of PVs in corpora is probably the frequency of PVs". *Absolute frequency* is the raw count of "all occurrences of a particular word in a corpus", in contrast to *relative frequency*, which will be used to refer to normalised data (Brezina, 2018:42-43) (see §3.8.2.2 for an explanation of how normalisation will be applied in this research). *Relative frequency* can be thought of as the mean frequency, which indicates a sort of average, according to Brezina (2018:43).

It is necessary that a distinction be made between the base form, or lemma, of a PV, and all occurrences of that PV. Weisser (2016:149) describes *type* as "a representative instance/word form in a frequency list", and *token* as "each individual occurrence of a particular type". Brezina (2018:39) defines *type* as a "unique word form in a corpus", and *token* as "a single occurrence of a word form". Brezina (2018:39) clarifies this definition of *type* as meaning that words like *time* and *times* will be reported on separately. However, such a use of the term would not be appropriate under the present circumstances, because of the resulting inconsistency of considering PVs such as *take out* and *taking out* as separate PVs. Rather, Gardner and Davies (2007:345), in their research into frequently used PVs, grouped together "all inflectional forms of the same verb". The same process was followed by other researchers into PVs, such as Chen (2013b). Therefore, in this study, the term *type* will be used to identify the base form (lemma) of a PV, and the term *token* will refer to occurrences of all variations of the base form. Zhou and Wang (2024:4) consider the type / token ratio as a useful way of representing "diversity of PVs used by English learners". Thus, distinguishing between type and token is valuable.

3.6 Pilot study

The primary reason for conducting a pilot study is to test the assumption of the proposed research and to uncover any anomalies that might have a negative impact on it. The pilot study used a longitudinal corpus of L1 and L2 student texts to investigate PV use. Apart from assessing how L1 and L2 students use PVs in their writing, as well as comparing L1 and L2 PV use, the availability of a longitudinal corpus allowed for an exploration of changes in PV use over a three-year degree. In order to gain a thorough illustration of student PV use, the investigation included a comparison of PVs and their one-word synonyms.

The WITS corpus, one of two corpora to be used in the main study, was selected for this purpose. This corpus, created by Cooper (2016), is a longitudinal corpus of 8 424 209 tokens. Its substantial size allowed for comprehensive testing of the proposed research parameters. In addition, the longitudinal nature of the corpus meant that an additional aspect of PV use could be assessed, namely the effect that prolonged exposure to an academic environment might have on PV use.

The methodology used for the pilot study will be presented here, consisting of an overview of the participants, the preparation of the corpus, and the method used to assess PV use. As many of these aspects will be discussed in detail in the main study, descriptions here will necessarily be brief. The findings of the pilot study will also be discussed.

3.6.1 Participants

The participants are Psychology students who studied at the University of the Witwatersrand over a three-year period. The assignments that make up the corpus were gathered from these participants over three consecutive years. While Cooper (2016) was only interested in the students who completed their studies and submitted the requisite number of assignments over the three years, no such constraint was needed for the pilot study. Therefore, all student submissions of this period were useful to and included in the pilot study. Participant ages ranged from 17 to 34, and most were female (88%).

3.6.2 Preparation of the corpus

As the aim of the pilot study was to compare L1 and L2 PV use, the corpus needed to be separated into L1 and L2 student groups. The corpus also required tagging for parts of speech to ensure the correct extraction of *verb + particle* combinations by Wordsmith Tools. For the tagging, Sketch Engine (Kilgarriff, Rychlý, Smrž & Tugwell, 2004) was used, which is a corpus manager and text analysis software. A thirty-day free trial of this software was available, which was a sufficient time period to complete the tagging of the two subcorpora (L1 and L2).

3.6.3 Assessment of PV use

Wordsmith Tools was used to generate a list of all occurrences of PV use found in the corpus of L1 student writing for all three years, using the Concord function. The PV extraction process is discussed in detail in §3.7.2. The Concord function provides several kinds of information, the most useful being a concordance list, which gives examples of sentences containing the PVs in the selected corpus, and a cluster list, which shows "patterns of repeated phraseology" (Scott, 2022).

Once all valid PVs had been identified, only the ten most frequently used PVs were retained because of the limited time available for the completion of the pilot study, and also because it was deemed an adequate sample of the data from which to make deductions. The *Collins Online English Dictionary* (2021) was used to find one-word alternative verbs (ALTs) for each of the PVs on the list. For example, for the PV *carry out*, the dictionary had the following entry: "***carry out*** *1. To perform or cause to be implemented 2. To bring to completion; accomplish*". In this case, *perform* and *accomplish* were used as ALT options, as *cause to be implemented* and *bring to completion* did not deliver results.

Occurrences of the use of each of the identified ALTS were recorded, using Wordsmith Tools. The same process was repeated for the L2 student writing. Using the information on PV use generated thus far, L2 PV use was then compared to L1 PV use. Examples of the results are provided in §3.7.4.

Finally, because the corpus contains longitudinal data, first-year L1 PV use was compared to third-year L1 PV use to determine if prolonged exposure to an academic milieu had influenced PV use. For the same reason, first-year L2 PV use was compared to third-year L2 PV use. Third-year PV use for L1 students was subsequently contrasted to the third-year PV use for L2 students to determine if the PV use of these two groups had become similar over the three years of study.

3.6.4 Findings

The ten most frequently used PVs in L1 student writing are given in Table 3.1, showing absolute as well as relative frequencies. The relative frequencies resulted from normalisation being applied to the absolute frequencies (see §3.8.2.2), in order to facilitate comparison between the L1 and L2 groups. Relative frequencies were rounded up to the nearest integer.

Table 3.1: Ten most frequently used PVs in WITS L1 student writing

PV	***Absolute frequency***	***Relative Frequency****
GROW UP	339	67
MAKE UP	253	50
CARRY OUT	195	38
BRING UP	83	16
ACT OUT	67	13
POINT OUT	62	12
SET OUT	55	11
PICK UP	47	9
GO ON	40	8
PASS DOWN	39	8

*Per million words

According to Biber *et al.* (1999), PV use is noteworthy for frequencies of 40 or more occurrences per million words. Biber *et al.* (1999) based this benchmark on their work with the *Longman Spoken and Written English Corpus* (LSWE), which has a token count of 40 026 000. This is substantially higher than that of the WITS corpus (8 424 209 total token count). However, the LSWE is itself made up of subcorpora that vary in size from 2 480 800 to 6 904 800 tokens, to which the standard of 40

occurrences per million words is consistently applied. Hence, it seemed reasonable to retain the benchmark here. (This benchmark was not applied in the main study, since the low PV frequencies of the other subcorpora used in the study made its use unsuitable, as will be seen in §3.7.2.1).

Accordingly, for the L1 group, three PVs indicate noteworthy frequencies of use (indicated in light grey). These are *grow up* (67 occurrences per million words), *make up* (50 occurrences per million words) and *carry out* (38 occurrences per million words), although, in the case of the PV *carry out* the frequency was rounded up to fall within the category considered noteworthy. When compiling the *Longman grammar of spoken and written English*, Biber *et al.* (1999:39) reported rounded figures as they found that "[S]everal factors can cause minor fluctuations in the frequency counts computed for lexico-grammatical features." While rounded frequencies were not used as a general rule in the pilot study, and given the arguments put forth by Biber *et al.* (1999:39), it was thought to be beneficial in a situation where the count proved to be close to a higher category, as with *carry out*.

The same process was repeated in order to extract and analyse the ten most frequently used PVs in WITS L2 student writing (not included here for the sake of space), after which the two groups of data were compared. The results of the investigations into WITS L1 and L2 PV use are summarised in Table 3.2 below, and represent total PV use over the course of the degree. The ten most frequently used PVs for both the L1 and L2 groups are given, along with their relative frequencies. Information about ALTS is also included for the L1 and L2 PVs represented in the table.

Table 3.2: Comparison of L1 and L2 PV and ALT use

L1				***L2***			
PV	***Relative frequency****	***ALT***	***Relative frequency****	***PV***	***Relative frequency****	***ALT***	***Relative frequency****
GROW UP	67	*mature*	17	GROW UP	91	*mature*	7
MAKE UP	50	invent	2	MAKE UP	50	invent	1
CARRY OUT	38	perform	188	CARRY OUT	38	perform	244
		accomplish	12			execute	5
BRING UP	16	raise	49	END UP	34	arrive	6
						land	0
ACT OUT	13	*demonstrate*	0	FIND OUT	17	determine	120
		illustrate	0				
POINT OUT	12	indicate	105	COME UP	16	arise	45
SET OUT	11	present	21	GO ON	13	continue	127
		arrange	7			persevere	1
		display	0				
PICK UP	9	gain	110	POINT OUT	11	indicate	75
		grasp	11			specify	5
GO ON	8	continue	106	PICK UP	9	gain	91
						grasp	12
PASS DOWN	8	bequeath	0	TURN OUT	9	become	314
		leave	57				
		transfer	27				
		bestow	2				
		donate	0				
Total	232		638		288		1053

*Per million words

For the L2 group, the use of the PVs *grow up, make up* and *carry out* was shown to be noteworthy, with the relative frequency for *carry out* being rounded up as for L1 PV use. Interestingly, not only

were the first three most frequently used PVs in WITS student writing (*grow up*, *make up* and *carry out*) the same for both L1 and L2 students, but two of them also had the same relative frequency (*make up* and *carry out*). This peculiarity could be the consequence of shared assignment topics predisposing students to the use of particular PVs (Immelman & Cooper, 2023). Comparing the total relative frequencies of the ten most frequent PVs for L1 and L2 students indicates that PV occurrences are 1.24 times more prevalent among L2 students than L1 students. Thus, it seems that L2 students are more likely to use PVs than are L1 students.

As indicated, ALTS are included in Table 3.2, as well as their relative frequencies. The process of establishing the suitability of the ALTS consisted of sourcing possible synonyms from the PV entries in dictionaries and, subsequently, extracting synonym frequencies from the relevant corpora by using Wordsmith Tools. This process produced concordance lines which provided context for the use of the ALT. Consequently, it was possible to establish that some synonyms were not appropriate ALTs, given the context. For example, for the PV *act out*, the synonym *demonstrate* produced the sentence *'…during a short setback. Note that injury of the prefrontal cortex* ***demonstrates*** *a positive correlation with performance in the multitasking condition…'*. However, *demonstrates* cannot be replaced by *acts out* in this context, and the use was thus flagged as unsuitable. Likewise, the synonym *mature*, as used in the sentence *'…increases each year in schoolchildren as their cognitive processes* ***mature****…'*, is not an appropriate ALT for the PV *grow up.* Such synonyms are indicated in italics in the table; their occurrences were not included when comparing PV and ALT use. As there are only four instances (*grow up*, *make up*, *act out*, and *pass down*) where L1 students preferred the use of the PV over that of the ALT (indicated in light grey), ALTs could be said to have been preferred 60% of the time. The investigation into the use of ALTs identified in L2 writing suggested that ALT occurrences are 1.65 times more prevalent among L2 students than L1 students.

An initial conclusion drawn from these findings is that the L2 student preference for ALTs rather than PVs aligns with previous research (Dagut & Laufer, 1985; Hulstijn & Marchena, 1989; Liao & Fukuya,

2004). On the other hand, previous research also showed a preference for PVs over ALTs by L1 students, which was not observed in the pilot study. It has been suggested that this might be a result of an emphasis on formality within South African academic institutions (Immelman & Cooper, 2023).

Table 3.3 presents the changes in L1 and L2 PV use over the three years of an undergraduate degree. In this case, PV totals were extracted for the first year (2011) and last year (2013) of study. This investigation looked into the possibility of alignment of L2 PV use with L1 PV use after extended exposure to an academic environment.

Table 3.3: L1 and L2 student PV use during a three-year undergraduate degree

L1 2011		***L1 2013***		***L2 2011***		***L2 2013***	
PV	***Relative frequency****	***PV***	***Relative frequency****	***PV***	***Relative frequency****	***PV***	***Relative frequency****
GROW UP	102	CARRY OUT	54	GROW UP	151	CARRY OUT	59
MAKE UP	62	MAKE UP	39	MAKE UP	52	GROW UP	26
CARRY OUT	41	GROW UP	37	CARRY OUT	28	MAKE UP	26
BRING UP	25	SET OUT	21	COME UP	27	END UP	25
FIND OUT	13	PASS OWN	16	END UP	20	FIND OUT	15
ACT OUT	10	TAKE UP	14	FIND OUT	18	SET UP	15
SET OUT	10	ACT OUT	13	BRING UP	13	GO ON	12
POINT OUT	7	PLAY OUT	11	BEAT UP	7	POINT OUT	11
COME UP	6	BRING UP	10	SUM UP	7	ACT OUT	9
COME OUT	6	POINT OUT	9	POINT OUT	7	SET OUT	9
Total	**282**		**225**		**330**		**207**

*Per million words

The first interesting observation is that the top three PVs across all three categories (L1 2011, L1 2013, L2 2011 and L2 2013) are the same: *grow up*, *make up* and *carry out*. When this pattern of same PV use between the L1 2011 and L2 2011 groups was first noticed, it was thought that this might be the result of the assignment topics that had been set. However, as a variety of different topics would have been covered by the third year, and as a great deal of reading would have been required from students during this period, it is somewhat unexpected to find the same PVs still emerging as the most

frequently used. Hence, it seems likely that there is some other reason for the popularity of the PVs, such as a high occurrence of these PVs in the popular media, resulting in their becoming entrenched in the students' language. On the other hand, frequency of use (per million words) varies in interesting ways over the three-year period. For both the L1 and L2 groups, the use of the PVs *grow up* and *make up* drops substantially (though still retaining a top three position on the table, illustrating the general drop in PV use by the groups), while the use of *carry out* increases. This very specific pattern does indeed seem to support the original idea that students were swayed in their PV use by assignment topics. However, further investigation is necessary to establish whether there is a correlation between the use of specific PVs and the subject matter about which students are writing.

The totals of the ten most frequently used PVs per student group for 2011 and 2013 were then considered. It was observed that the first-year L2 student group showed the highest PV use (with a relative frequency of 330) for the three-year period, followed by the first-year L1 student group (with a relative frequency of 282). This observation is interesting in that it contradicts patterns of PV use observed in other research, where L2 students, in contrast to L1 students, were generally found to avoid PV use (Hulstijn & Marchena, 1989; Liao & Fukuya, 2004). Furthermore, the third-year L2 student group demonstrated the lowest PV use (with a relative frequency of 207). This is again at variance with previous research, which showed L2 students gaining confidence in PV use after increased exposure (Hulstijn & Marchena, 1989; Liao & Fukuya, 2004).

Based on the findings above, it was concluded that L2 students in their first year of study (as represented in the WITS corpus) are more inclined to use PVs in their academic work than do L1 students. However, this inclination declines with increased academic exposure, which is a pattern that is also evident in L1 student writing. These observations seem to underscore the idea that students, both L1 and L2, move towards a more formal writing style in an academic environment.

The pilot study was useful in establishing the parameters that would be used in the proposed research. Nevertheless, a few factors needed to be modified to enhance the results of the main study. For

example, for the pilot study, the minimum frequency (that is, the minimum number of times that a cluster had to appear in the corpus to be reported on) was set to five. After investigation (see §3.8.2.1.1), this was found to substantially affect results because of the number of low-occurrence PVs that were not being included. In addition, the method of PV use authentication needed to be improved by incorporating a more meticulous checking of the concordance lines, which would, furthermore, enhance error reporting.

3.7 Main study

In this section, the methodology to be used in the main study will be discussed. This will include an explanation of the two corpora that form the basis of the research, and a detailed discussion of the data collection and handling process. Data collection and processing incorporates both quantitative (PV data extraction and preparation) and qualitative (PV verification) methods. The process to be used for error identification will also be discussed, as well as the method of obtaining synonyms for the most frequently used PVs that had been extracted from the corpora.

3.7.1 Corpora

Sufficient data should be used in research to ensure that the patterns of use that are detected are authentic, and that legitimate conclusions are drawn. Hence, with the arrival of large, machine-readable corpora, researchers have been able to "yield comprehensive pictures of students' production of PVs in their real writing" (Chen, 2013b:91). In line with Chen's (2013b) viewpoint, McEnery and Hardie (2012:2) assert that "a corpus is best used to answer a research question which it is well composed to address". This research study makes use of two sizeable corpora, the South African Multilingual Learner Corpus of Academic Texts corpus (SAMuLCAT) and the University of the Witwatersrand corpus (WITS), both of which include L1 and L2 undergraduate student texts. The SAMuLCAT corpus comprises samples of student writing from various South African universities, whereas the WITS corpus comprises samples from one South African university only. A further difference between the two corpora is that the SAMuLCAT corpus does not contain longitudinal data,

and so does not allow for investigation into changes in student PV use over the course of a degree, as was possible in the pilot study which used the longitudinal WITS corpus. Moreover, new data are added to the SAMuLCAT corpus at a fairly consistent rate, with the result that corpus data are reasonably current. The corpus can therefore be said to be fairly representative of current student writing. On the other hand, the WITS corpus is closed, having been finalised in 2016, meaning that the data are not as current as that of the SAMuLCAT corpus (see also §3.7).

The corpora include student and institutional metadata, which makes it possible to differentiate between L1 and L2 speakers. This distinction is essential to the investigation of patterns of PV use by L1 and L2 students, as well as the identification of differences in PV use between the two groups. The allocation of each student as either a mother-tongue or second-language speaker of English is based on metadata provided by participating students when giving permission for the use of their data.

Wordsmith Tools will be used to investigate patterns of PV use, which will require that the words in the two corpora are labelled, or tagged, for parts of speech. The SAMuLCAT corpus is made available to researchers in a tagged format, having been tagged for parts of speech using NLP4J (The Natural Language Processing for JVM languages project), but the WITS corpus is untagged in its original form. However, because this corpus was used in the pilot study, it was tagged at that stage by the researcher using Sketch Engine, as this software was accessible and easy to use. In neither case was quality control conducted on the automated tagging, and minor errors are to be expected due to, among other things, the complicated nature of parts-of-speech identification (Biber *et al.*, 1999:39). Since an aspect of the study will include qualitative research, human error could also be a factor. As Biber *et al.* (1999:39) assert, "when working with large corpus files, there will always be some variability due to error, depending on occasional human mistakes or inaccuracy of methodological tools". Brezina (2018:262) further states that "[e]ven well-established corpora such as the BNC include errors, inconsistencies and possible bias."

Gardner and Davies (2007:342) acknowledge the potential problems inherent in trusting electronically tagged corpora to correctly identify relevant parts of speech, such as the adverbial particle so essential to the definition of the PV. Therefore, they suggest that PVs are probably underreported in electronic tagging (Gardner & Davies, 2007:342). As examples of where such errors might occur, they provide the sentences '*you can* ***turn in*** *for the night*' (*verb + adverbial particle*) as opposed to '*the police told me to* ***turn in the opposite direction***' (*verb + the prepositional phrase*). Nevertheless, as they themselves acknowledge, such errors are probably negligible according to the report produced by BNC (the corpus used in their research) which indicates a classification error rate of 1.58% for adverbial particles and 0.59% for prepositions.

The disadvantages of corpus research need to be acknowledged. For instance, in their study on the meaning senses of the most frequently used PVs, Liu and Myers (2020:421) recognise that "the findings from any corpus study are always limited by the data in the corpus", since not all language used is necessarily represented in the corpus. The fact that the speech subcorpus of the COCA contains primarily broadcasting language, which might differ substantially from conversational American English is a case in point (Liu & Myers, 2020:421). In this regard, Baker (2006:17) also points out that a speech corpus does not necessarily incorporate all the elements that accompany speech, such as body language. Zarifi and Mukundan (2014:53) remarks that using only a limited sample of a corpus could "put the representativeness of the corpus into question". Furthermore, Chen (2013a:419) points out the problem of finding appropriate corpora for the comparison of L1 and L2 writing because of the possible influence of extralinguistic factors (such as differences in sources, and time allowed for the completion of a task) on the results. The effect of the variety of English (for example, British versus American English) in the native corpus to be used in the comparison is another factor to be considered (Chen, 2013a:419).

The present study is in the fortunate position that the corpora that are involved represent a fairly homogenous group of South African students. This means that varieties of English are not at issue, as

the texts incorporated into the corpora are largely written in South African English. Moreover, L1 and L2 comparisons are conducted within a corpus, where the same conditions existed for both groups for the completion of assignments. It should be noted that the assignments contained in both corpora cover a range of genres, topics and subjects (see §3.8.1.1.3 and §3.8.1.2.1). It was, therefore, not possible to determine whether the assignment topics had an influence on the choice of PVs used, as the level of detail required for such analysis was beyond the scope of this research.

3.7.1.1 *SAMuLCAT*

Information about the SAMuLCAT corpus is provided here. This includes a description of the corpus, the reasons for its inclusion in this study, and an overview of the participants in the corpus. The method used to prepare the corpus for use in this study will then be explained.

3.7.1.1.1 Description of corpus and data collection process

The SAMuLCAT corpus was set up by the Inter-institutional Centre for Language Development and Assessment (ICELDA) in partnership with the South African Centre for Digital Language Resources (SADiLaR) (Carstens & Eiselen, 2019:67). South African higher education institutions that belong to the ICELDA network contribute to the corpus on an ongoing basis, which means that the corpus is ever expanding, making it an excellent resource for researchers. As several institutions make contributions, this also means that the corpus includes a variety of genres, lengths of texts, levels of student competence and fields of study. While the collected writing ranges from first-year to post-graduate level, it was not collected from the same set of participants. The corpus is, therefore, not longitudinal. The corpus incorporates texts in various South African languages, although only the English texts were relevant to the present study. The data of two institutions included in the corpus were available to the researcher. These were North-West University (NWU) and the University of Pretoria (UP). The corpus also includes data for Stellenbosch University (US), but this institution was eventually excluded from the research, for reasons given in §3.8.1.1.5.

The corpus was annotated for parts of speech, although the taggers used for South African indigenous languages differs from that used for English. For the English data, *Natural Language Processing for JVM languages (NLP4J)* had been used for parts of speech tagging. Further processing was required to prepare the data for correct PV extraction using Wordsmith Tools. A more detailed discussion of the data processing is given in §3.8.1.1.4.

The SAMuLCAT corpus is open-source and in the public domain. The necessary ethical clearance for the collection of the data contained in the corpus is obtained from the individual institutions, as well as the informed consent of all the contributing students. The data collection process produces two kinds of metadata: firstly, lecturers or relevant staff members provide information related to the particular tasks or assignments; and, secondly, students provide their own biographical data. The metadata were used to separate students into L1 and L2 speakers for this study. Even though the corpus is open source, the use thereof requires ethical clearance from one's own institution. Once ethical clearance has been obtained, permission for the use of the corpus must then also be requested from SADiLaR. These procedures were duly performed by the researcher and permission for the use of the corpus was granted. (Refer to Appendices A, B and C for the signed letters of permission.)

3.7.1.1.2 Reason for inclusion in the study

The SAMuLCAT corpus was suitable for inclusion in this study for several reasons. Because of its size, the conclusions drawn about the patterns of PV use that were encountered could be claimed to be legitimate. Furthermore, the corpus comprises student texts from several South African higher educational institutions, making it possible to compare PV use across institutions.

3.7.1.1.3 Participants

Participant profiles differ slightly for the two subcorpora within the larger SAMuLCAT corpus. Participants in the NWU corpus mainly consist of first-year students, although the texts of second- and third-year students, and even some post-graduate students, are also included. The students fall within the 16 to 40 age range. The writings submitted are Academic Literacy tasks and assignments within

the Humanities and Social Sciences Faculty. To study Social Sciences, an Admission Point Score (APS) of 22 is required. In addition, students must achieve a matric English first-language mark of at least 50%, or a matric English first-additional language mark of at least 60%.

The UP corpus participants are first-, second-, and third year students, with an age range of between 16 and 40. Academic Literacy tasks and assignments within the Health Sciences, Humanities and Social Sciences, and Economic and Management Sciences subject fields were submitted. Requirements for admission to Health Sciences, depending on the chosen course, are AP scores that range from at least 25 to 35, and a matriculation English mark of at least 50% or 60%. Students who wish to study in the Humanities and Social Sciences field require, depending on the chosen course, a minimum AP score that ranges from 26 to 34, and an English mark of at least 50% or 60%. Finally, Economic and Management Sciences students will require, again depending on the course, AP scores of 26 with an English mark of at least 50%, up to AP scores of at least 34, with an English mark of at least 60%.

In summary, the student composition of the NWU and UP subcorpora appears to be similar, in that the level of English within each subcorpus can be expected to vary according to selection criteria. For this reason, it will be difficult to assess the influence of selection criteria on English competence, and PV use, in particular.

3.7.1.1.4 Preparation of the NWU and UP subcorpora

The two SAMuLCAT subcorpora needed to be prepared before data extraction could take place. The same preparation process, as described here, was used for both. They were received in XML format and, as mentioned previously, had already been tagged for parts of speech using the *Natural Language Processing for JVM languag*es (NLP4J) platform, available from an open-source library (obtainable at https://emorynlp.github.io/nlp4j/). The following is an example of how a particular word, *rhinos*, lemmatised and tagged according to its part of speech, appears in the corpus:

<w lemma="rhino" type="NNS">Rhinos</w>

The “w” symbol indicates the start of a word, and “</w>” the end of a word. Two pieces of information are provided per word: the base form, or lemma, of the word, and the part of speech (“type”) to which it belongs. The part of speech codes are those used by NLP4J: in this case, “NNS” indicates that the word is a plural noun.

After initial tests, it was found that the lemma identifier would need to be removed for Wordsmith Tools to effectively identify PVs (*verb proper + particle*) in the corpus. As no method could be found to automatically remove the lemma identifier, it was done manually. Fortunately, this only needed to be done for verbs and particles, the constituents of PVs, and thus the only part of speech that needed to be identified by Wordsmith Tools.

This proved to be a worthwhile process as it provided insight into what Wordsmith Tools would encounter and report on. For example, the number of words that might separate the verb and particle of a PV was regarded by the Wordsmith Tools program as three at most, as in ‘*The children **make** the fantasy stories **up***’. However, the manual perusal of the corpus indicated that many more words might be used to separate the verb and particle, such as in *‘…**dragging** the social image of the individual **down**…’*, where six words separate the verb and particle, and in *‘…**turn** the GPS system connected to the consumer side of the device **off**…’*, where eleven words separate the verb and particle. Knowledge of such PVs would have been missed had the researcher not been made aware of their existence as a result of having to conduct a manual search for PVs. Eventually, for practical reasons, the decision was made to extract only PVs that were separated by three words (see §3.8.2.1.1).

There were also instances where the tagger had incorrectly identified the verb of the PV as a noun (as in *‘…**end** [noun] **up**…’*), requiring manual adjustment to reflect the correct part of speech (in this case, *‘…**end** [verb] **up**…’*). Such errors were sometimes the result of a grammatical error on the part of a student, such as where a missing apostrophe caused the tagger to tag **’its closing down’* as *pronoun + noun + particle*, instead of *pronoun + third person singular verb + present participle + particle*, as

indicated by the context of the sentence. Again, these PVs would have been missed by Wordsmith Tools had the manual process not alerted the researcher to their existence.

One other issue necessitated further modification of the corpus before analysis could start. The underlying assumption for the proposed research was based on a comparison of PV use between L1 and L2 speakers. The SAMuLCAT corpus does not make this distinction. Rather, texts are grouped by assignment. Therefore, it was the task of the researcher to separate student records into L1 and L2 files. SAMuLCAT metadata provide various language markers for each student (*father's home language, mother's home language, school home language* and *school language*), but do not include a marker indicating the student's home language. Consequently, the *school home language* marker was used as the pertinent language marker, as it indicates the language that students specified as their home language while at school.

3.7.1.1.5 A note on the University of Stellenbosch subcorpus

The University of Stellenbosch (US) is a contributor to the SAMuLCAT corpus and, thus, forms one of its subcorpora. This subcorpus was not available in XML but rather in EXCEL format and, therefore, needed to be processed separately from the rest of the subcorpora (NWU and UP). As the same metadata had been collected for this subcorpus, it was separated into L1 and L2 subcorpora using the same language marker criteria as for the NWU and UP subcorpora (s*chool home language*). At this point, the subcorpus could be investigated using Wordsmith Tools in the same way as for all the other SAMuLCAT subcorpora. However, with 408 871 tokens (220 063 L1 tokens and 188 808 L2 tokens), this subcorpus proved to be too small for inclusion in the research, as normalisation produced values that were not viable. Therefore, it was decided to exclude the US subcorpus.

3.7.1.2 *WITS corpus*

Whereas the SAMuLCAT corpus consists of data from several institutions, the WITS corpus contains data from only one institution. This corpus was briefly discussed in the section describing the pilot study and will be expanded on here. Information about the corpus is taken from Cooper (2016).

There are a few provisos that should be noted regarding this corpus. The Wits corpus was created between 2011 and 2013. It is therefore already almost ten years old and might not be representative of current PV use. Replication of the study with more current data might provide interesting insights into changes in PV use, possibly due to an increase in the use of social media. However, it is not foreseen by the researcher that PV use would have greatly altered during this time, except that newly coined PVs might have come into use. A further point to be noted is that student writing is only based on Psychology. The extent to which these factors had an influence on the results are not addressed in this research, but might be interesting topics for future research.

3.7.1.2.1 Description

The WITS corpus, a longitudinal corpus consisting of 8 424 209 tokens, was created by Cooper (2016) over a three-year period. Data, in the form of written assignments, were collected from first-, second- and third-year students doing a Psychology degree at the University of the Witwatersrand. Psychology is a popular course for which a high number of students was expected to register, and also to continue through to the third year of study. To be included in the study that Cooper (2016) was conducting, students not only had to complete the three-year course, but also to submit a specific number of assignments per year. The present study was focussed on the PV use of L1 and L2 students generally rather than following the progress of specific students, and hence there was no restriction on the assignments used. The metadata of participating students were used in the allocation of students as L1 or L2 speakers according to the file identifiers provided in Cooper's (2016) study.

3.7.1.2.2 Reason for inclusion in the study

The size of the WITS corpus makes it a substantial corpus, well suited to the requirements of corpus research. Because it is drawn from a higher education institution not represented in the SAMuLCAT corpus, it adds diversity to the research. It also meets the necessary requirements for this study, in that the writing of both L1 and L2 students are available, in the form of academic essays within the field of Psychology.

3.7.1.2.3 Participants

The corpus comprises assignments completed by students enrolled for Psychology at the University of the Witwatersrand, from 2011 to 2013. As Cooper's (2016) study was longitudinal, in that it covered student progress over a three-year degree, only the assignments of those students who completed the course, and complied with certain requirements, were selected for use in the study. Therefore, first-year assignments were selected retrospectively, based on those who had continued on to their second year. At the end of the third year, only those students who had met all the criteria over the three years were included in the final analysis. In the end, of the 782 students that originally registered for the course, the data of 87 students were collected for the 2011 subcorpus, of 208 students for 2012, and of 160 for 2013. However, all collected assignments were used in the current study, as individual student progress over the three years was not under consideration in this case.

Cooper (2016) notes the possibility of research involving WITS students being somewhat skewed and provides the following reasons for this. WITS is a prestigious higher education institution which attracts high-performing students with above-average matriculation results. The WITS student thus "presents a profile of the more privileged and less disadvantaged students in South African society" (Cooper, 2016:100). In addition, the students wishing to register for Psychology need to have achieved at least 60% for English in matric. In general, therefore, WITS students can be expected to have a fairly competent level of English on the whole.

3.7.1.2.4 Preparation of the corpus

Because of a different research focus, the corpus needed to be adapted for use in this study. (This process was carried out before the pilot study was conducted.) Firstly, the data had to be separated into L1 and L2 subcorpora, for which purpose the metadata were used. In the WITS corpus, the metadata are stored in a file that is separate from student texts, although a unique code links a student's writing and demographic information. Hence, the metadata language identifier *1st/2nd language* was used to identify and separate L1 and L2 texts.

The two subcorpora then needed to be tagged for parts of speech, so that PVs could be identified (*verb + particle* combinations) when using Wordsmith Tools. This was done using *Sketch Engine*. However, the format in which *Sketch Engine* delivers tagging was not suitable for use by the Wordsmith Tools software. An example of its tagging format is given here:

Many	JJ	many-j
people	NNS	people-n
in	IN	in-i
South	NP	South-n
Africa	NP	Africa-n

where *JJ* indicates an adjective, *NNS* indicates a plural noun, *IN* indicates a preposition or subordinating conjunction, and *NP* indicates a proper noun. Fortunately, Wordsmith Tools has a *Text converter* function that can be used to convert the tagged subcorpora into the correct format, which meant that the text above appeared as follows after being converted:

<JJ>Many <NNS>people <IN>in <NP>South <NP>Africa

The data was then ready for data collection using Wordsmith Tools.

The fact that different tagging procedures were used for the two corpora needs to be addressed briefly. As discussed above, circumstances dictated the use of two tagging systems, but it was still necessary to ensure that the results were not affected by this decision. Therefore, the following test was conducted, using one of the HTML files provided by SAMuLCAT. The particular file was selected

simply for size (approximately 200 000 tokens), so that it provided sufficient data on which to base a decision. Most importantly, the file, selected from the NWU subcorpus and consisting of both L1 and L2 data, was available in both a tagged (NLP4J), as well as in an untagged (text) format. This made it possible to compare NLP4J tagging to Sketch Engine tagging, to establish whether differences in PV identification might adversely affect results.

The untagged file was tagged using Sketch Engine, and the further necessary processes were followed to prepare the file for data extraction using the Wordsmith Tools Concord function (as detailed above). The NLP4J file was then also submitted to Wordsmith Tools, so that two lists of PVs, resulting from identical files that only differed as far as the tagging platforms were concerned, were now available for comparison.

The same PVs were identified in both cases, except for three instances where the particle was identified as an adverb (in the case of the Sketch Engine tagger), and two instances where the particle was identified as a preposition (in the case of the NLP4J tagger). The result was that the NLP4J tagged version of the file produced a total of 29 legitimate PVs, while the Sketch Engine tagged version of the file produced 27 legitimate PVs. The file that had been tagged using the NLP4J tagging system, therefore, identified 3.4% more PVs. While it is not possible to make a categorical statement based on such a small sample, this outcome, nevertheless, suggests that a small percentage of additional PVs is more likely to have been identified in the SAMuLCAT corpus than in the WITS corpus. However, as the results show (see §4.3), the WITS corpus, tagged using Sketch Engine, consistently produced more PVs than the SAMuLCAT corpus, so that the differences between the tagging platforms cannot be said to have influenced the results.

The information in this section covered descriptions of the corpora to be used in the study, namely the SAMuLCAT corpus, comprising the NWU and UP subcorpora, and the WITS corpus. Reasons were provided for the inclusion of these corpora, and the exclusion of the US subcorpus. The preparation of the corpora for use in this study was described in detail.

3.7.2 Data collection and processing

In this section, the process used for data collection and processing will be discussed. Both the quantitative (the extraction of *verb + particle* combinations) and qualitative methods used in the study (the authentication of PVs, which includes the identification and documentation of errors) are described.

3.7.2.1 Verb + particle extraction *(quantitative)*

A PV is defined as a multi-verb combination of *verb + adverbial particle*, which may or may not be separated by another word or words. (For a full discussion of the definition, see §2.3 of the Literature review.) The Concord option of Wordsmith Tools was used to identify all *verb + particle* combinations. This resulted in the generation of PV frequency lists, providing quantitative data. (A concordance list of sentences that contain PVs was also generated, and will be used for qualitative analysis.) Corpora are particularly useful for such quantitative analyses (McEnery & Wilson, 2001:81).

3.7.2.1.1 Extraction parameters

Certain decisions needed to be made as to the parameters that would be used for the analysis of the data. For example, what would be considered a feasible number of PV cluster occurrences? On the one hand, one could argue that where only a small number of occurrences of a particular PV is found, considering the additional time and effort involved in processing such combinations, the PV is not noteworthy and does not require reporting. On the other hand, valuable information, such as the incorrect formation of PVs, might be missed if only noteworthy numbers of cluster occurrences are selected.

The most practical way to address this question was to test both scenarios by doing a concordance in which PVs that occur five times or more were reported (as in the pilot study, where the minimum frequency default setting of five had been used), as well as one where all occurrences of PVs were

reported. By comparing the two concordance cluster lists, it was possible to assess which version provided the most useful information.

The SAMuLCAT NWU L1 subcorpus was selected for this purpose, as, with 2 030 568 tokens, it was large enough to produce an authentic answer to the question under consideration. Firstly, using the Concord option, the cluster minimum frequency parameter was set to pick up only PVs that appear at least five times in the subcorpus (referred to hereafter as Test 1). All versions of the resulting 44 PVs were grouped according to the base form of the PV (thus, *grow up*, *growing up* and *grew up* would be grouped with the base form, or lemma, of the PV *grow up*). The lemmatisation produced 21 PV types (alphabetised in Table 3.4 below).

Table 3.4: SAMuLCAT NWU L1 PVs with five or more occurrences

PV	***Frequency***
BREAK IN	8
BRING ABOUT	25
BRING IN	5
BRING ON	5
CARRY OUT	25
COME UP	22
END UP	23
FIT IN	5
GO ON	15
GROW UP	20
KEEP UP	21
LEAVE OUT	12
LOOK UP	6
MAKE UP	23
OPEN UP	7
POINT OUT	5
SET OUT	17
SET UP	6
STAND OUT	10
TAKE ON	13
TURN OUT	5

The minimum frequency parameter was then set to one, so that all PV cluster occurrences in the subcorpus would be picked up, regardless of frequency (referred to hereafter as Test 2). This produced 688 unsorted PV clusters, a substantial increase on the 44 PV clusters found in Test 1. Table 3.5 gives a comparison of the distribution results of the two tests. The number of PVs with five or more occurrences (44) was, naturally, the same for both Test 1 and Test 2 (as the same subcorpus was used). Therefore, it is clear that the majority of PVs in the NWU L1 subcorpus occurred fewer than five times (688 – 44 = 644). In fact, most PVs seem to have occurred once only (476), and the next highest number of PVs (114) occurred twice only. Thirty-five (35) PV clusters occurred three times, and 19 occurred four times. Thus, the higher the frequency of occurrence, the lower the number of PVs at that frequency.

Table 3.5: Comparison of the results of Test 1 and Test 2

Frequency group	***Test 1***	***Test 2***
Five or more occurrences	44	44
Four occurrences	-	19
Three occurrences	-	35
Two occurrences	-	114
One occurrence	-	476
Total PVs	**44**	**688**

From this comparison, it is immediately clear that a substantial number of PVs are not identified if only PVs that occur five time or more are searched for. On the other hand, it is also necessary to determine whether the PVs that occur fewer than five times are worth reporting on.

An answer to this question emerged when the PVs were grouped according to the base form of the PV (PV type) as had been done for the PVs in Test 1. For example, for the PV *point out* (see Table 3.6), Test 1 gave a frequency of five. However, in Test 2, this PV has a frequency of 13. The reason for this higher count is that, in Test 2, in addition to the frequency of five for *point out*, there is also a frequency of three for *pointed out*, three for *pointing out*, and two for *points out*. These versions of the PV *point out* were not picked up by the first test because they occurred fewer than five times, which means

that only the present simple tense (plural) was reported on, and not the past simple, present continuous and present simple (singular) form. There is thus a difference of eight occurrences for the PV *point out* between Test 1 and Test 2.

Table 3.6: Frequencies for all clusters relating to the PV ***point out***

PV	***Frequency***
POINT OUT	5
POINTED OUT	3
POINTING OUT	3
POINTS OUT	2
Total	13

Another example of the impact that testing for all occurrences of PVs has, is illustrated by the PV *break down*. This PV was not picked up at all in Test 1, as the various versions that comprise the PV *break down* had low frequencies (*break down* = 1, *breaking down* = 1, *breaks down* = 1, and *broken down* = 3) (see Table 3.7 below). Nevertheless, the cumulative frequency of six is substantial enough to be reported. There are a notable number of PVs that fall within this category: *break down* (6), *bring out* (6), *figure out* (5), *find out* (5), *give out* (6), *hunt down* (5), *look out* (6), *reach out* (5), *take down* (5), and *take up* (5).

Table 3.7: Frequencies for all PV clusters relating to the PV ***break down***

PV	***Frequency***
BREAK DOWN	1
BREAKING DOWN	1
BREAKS DOWN	1
BROKEN DOWN	3
Total	**6**

There is also a third reason for including all PVs in the data analysis. In this case, the reason is qualitative rather than quantitative, as will be explained here. Even when the PVs had been categorised according to their base forms, there were 128 PVs that occurred once or twice only. These PVs, nevertheless, proved to be of interest, because of what they revealed about (in this case L1)

student PV use. For example, uses of such PVs as the casual *rev up*, and the more sophisticated *bring forth*, are interesting and worth reporting from a qualitative point of view.

We are now in a position to consider the question of whether reporting on all PVs found in a corpus is worth the considerably greater effort and time, or whether it should be considered sufficient to report only on higher frequency PVs. Upon considering the valuable information added to the analysis of the data by PVs which occur infrequently, as discussed in the previous paragraphs, it was decided to set the *Concord Minimum Frequency* parameter to pick up all occurrences of a PV in all further research in this study.

The *Words in Cluster* setting was also adjusted from its suggested setting. In the discussion on the preparation of the NWU and UP subcorpora (§3.7.1.1.4), separation of up to eleven words (thus, PVs of 13 words) were shown to exist in the corpus. Nevertheless, the decision had to be taken as to the expediency of picking up unnecessary combinations for the sake of a few outliers. Gardner and Davies (2007:345) limited their search to PVs of up to four words (i.e. with two words separating the verb and adverbial particle, such as in '***took*** *the nodules* ***back***') because of the infrequency of longer PVs. They found that there were relatively few PVs displaying longer separations, and they therefore did not search for longer strings, a practice confirmed by Liu and Myers (2020:411). In acknowledgement of these concerns while allowing for some leeway, the outer limit of the *Concord Minimum Frequency* parameter was set to five for this study.

3.7.2.1.2 Data preparation

The extracted clusters were saved as an EXCEL spreadsheet in preparation for further manipulation. The clusters were sorted alphabetically, and all forms of the PV type were grouped together. This is in keeping with general practice, according to McEnery and Wilson (2001:82), who state that "various forms of the same lexeme may be lemmatised before a frequency count is made: for instance, *loved*, *loving* and *loves* might all be considered to be instances of the lexeme LOVE" (italicised and capitalised as in original).

Table 3.8 presents an Excel sheet that was produced in this way, showing an example of one complete list of multi-word verbs that were found in the subcorpus for the verb *act* (*act out, act up* and *act in*). It illustrates how all versions of, for example, *act out* have been combined (*act out, acting out* and *acts out*). While the combination *acted out* could also have been expected to be found here, no such examples were found in this instance. By working through the list in this way, cases where clusters needed to be repositioned became evident, as in the case of *grew up* which needed to be grouped with *grow up*. (Subsequently, the *act in* cluster was discarded, not being a PV. This is discussed in §3.8.2.2.)

Table 3.8: Example of alphabetised cluster list in EXCEL

PV	***Cluster***
ACT OUT	
	ACT OUT
	ACT OUT AND
	ACT OUT IN
	ACT OUT MAKE
	ACTING OUT
	ACTING OUT DUE
	ACTS OUT
	ACTS OUT AGGRESSIVELY
	ACTS OUT BY
	ACTS OUT THIS
	ACTS OUT WHAT
ACT UP	
	ACT UP
	ACT UP IN
	ACTING UP
	ACTING UP ON
ACT IN	
	ACT IN
	ACTING IN
	ACTING IN UNACCEPTABLE

In this section, the process of producing a concordance and cluster list using appropriate extraction parameters was explained. In addition, the manipulation of the data contained in the cluster list was discussed. At this stage, the next step in the process is the qualitative verification of the combinations.

3.7.2.2 PV verification *(qualitative)*

Up to this point in the processing of clusters, the multi-word verb combinations had yet to be validated as PVs. Intuition cannot be relied on in deciding whether a multi-word verb is a valid PV, another kind of multi-word combination, or an incorrectly used PV. Close examination is required, which includes the checking of each combination against PV dictionaries, and the scanning of the relevant concordance lines. Chen (2013b:93) similarly found it necessary to separate PVs from other multi-word combinations as "Wordsmith Tools included not only PVs but also verb + particle free combinations or verb + prepositional phrases since many particles function as both adverbs and prepositions". She also included the use of dictionaries in her study to verify legitimate PVs.

Several dictionaries were used for this task for the following reasons. Because of the prolific nature of PVs, not all PV dictionaries contain a complete list of existent PVs. Furthermore, while basic PVs do appear in most of these dictionaries, there is some variation as far as lesser used PVs are concerned, which means that the most recently published PV dictionary is not necessarily the most inclusive. For these reasons, two PV dictionaries were used initially, namely the *Cambridge Phrasal Verbs Dictionary* (2006), and the *Collins COBUILD Phrasal Verbs Dictionary* (2013). Both of these were sourced online via the *Internet Archive* (2014), as the PV dictionaries available in the University library were much older, the most recent publication being the *Cambridge international dictionary of phrasal verbs* (1997). It subsequently proved useful to also consult three online dictionaries in order to identify recently established PVs. These online dictionaries are listed below:

a. *Collins Online English Dictionary* (2021)
b. *Merriam-Webster Online Dictionary* (2024)
c. *Oxford Learner's Dictionaries* (2024).

Most dictionaries, and certainly all of the dictionaries mentioned above, use the alternative definition for a PV of *verb + adverbial particle or preposition*, as discussed in the literature review. It was therefore, in some cases, necessary to apply some of the checks described in the literature review to

further validate a PV, such as the particle placement test, the unacceptable fronting in wh-questions, and the unacceptable fronting in relative clauses. For example, in the sentence '*The sunlight **reflects off** the water*'[SI], the two parts of the *reflect off* combination cannot be separated (*'*The sunlight **reflects** the water **off***'), a wh-question is allowable ('***Off** what does the sunlight **reflect***?'), as a relative clause is acceptable ('*The water **off** which the sunlight **reflected***'). According to these tests, therefore, *reflects off* is not a PV. On the other hand, *end off* (as in '*She finally **ended off** the long letter*'[SI]) allows for the placing of the particle after the direct object ('*She finally **ended** the long letter **off***'), but does not allow fronting in wh-questions (*'***Off** what did she finally **end** the long letter?*'), or the formation of a relative clause (*'*The long letter **off** which she finally **ended***'), meaning that *end off* can be confirmed as a PV.

Qualitative analysis is useful for exploring the quality of the PVs that students produce (Zhou & Wang, 2024:5). Once a PV had been validated against a dictionary entry, its use had to be checked in the context in which it appeared in the text. For this reason, the concordance lines, which had been saved as an EXCEL sheet, needed to be scanned manually. In order to do this, the EXCEL "find" function was used. For example, if the PV *cover up* was the PV being validated, the search string *cover <w type="RP">up* would have delivered results such as those shown in Table 3.9 below. All the concordance lines display correct use of the PV, except for line 6 ('*Fake news was known type="VBZ">has a **cover** <w type="RP">**up** <w type="IN">for <w type="VBG">being…*'). *Cover up* is a legitimate PV, but in the context of this concordance line it has been used as a noun, and thus it cannot be added to the number of legitimate uses of this PV. Thus, reading through the concordance lines from which the PV were derived was essential in confirming that the PV fitted the criteria identified in §2.3.2. This was the final proof of the validity of a PV. The processing of the different kinds of PV errors is discussed in §3.8.2.6.

Table 3.9: Example of concordance list in EXCEL for the PV cover up
Concordance*
or in that of his party , tries to type="VB"*>cover <w type="RP">up a financial scandal but can not do
can also be used to type="VB">cover <w type="RP">up other news that government organizations m
individuals engaged with debasement appear to type="VB">cover <w type="RP">up by accusing others
they <w type="VBP">end <w type="RP">up trying to type="VB">cover <w type="RP">up their tracks
that they have someone who is going to type="VB">cover <w type="RP">up for them should it happen
Fake news was known type="VBZ">has a cover <w type="RP">up <w type="IN">for <w type="VBG">being
every time you are late they have to type="VB">cover <w type="RP">up <w type="IN">for you
being involved in criminal activities to type="VB">cover <w type="RP">up the need for a father and to

*Key to part of speech symbols: VB=verb base form, VBZ=verb, third person singular present tense, IN=preposition or subordinating conjunction, RP=particle

As illustrated in Table 3.10 below, the base form of each valid PV was subsequently indicated in green, and was followed by all its iterations. Errors were indicated in yellow (followed by all iterations), with a short note indicating why the combination was not acceptable. (The colour coding facilitated the summarising of information by PV type at a later stage.) Each iteration of a PV (as in *act out, acting out* and *acts out*) included the number of occurrences, or frequency, encountered for that iteration. Therefore, *act out* occurred four times, *acting out* once, and *acts out* four times. As a perusal of the concordance lines featuring these PVs indicated no errors, the total overall frequency for the PV *act out* is therefore nine, as indicated in the *Total* column. If there had been any errors, they would have been subtracted from that total.

Table 3.10: Example of a cluster list in EXCEL indicating valid PVs

PV	*Cluster*	*Frequency*	*Total frequency per base form*
ACT OUT			
	ACT OUT	4	9
	ACTING OUT	1	
	ACTS OUT	4	
ACT UP			
	ACT UP	1	2
	ACTING UP	1	
	ACT IN **not a PV**	1	
	ACTING IN	1	

Acting in was highlighted in yellow, indicating that it was not a PV. (It is, in fact, a PrepV.) The occurrences for this combination were, therefore, not totalled. All issues thus indicated as being some other form of multi-word verb, be it a PrepV, a PrepP, or a Phrasal-Prep, were removed from the final list. All errors (see §3.8.2.6) were likewise removed and reported on separately.

Once the cluster list consisted solely of valid PVs, all extraneous information was deleted from the list so that only the base form of the PVs, each with its relevant total frequency, remained (as shown in Table 3.11).

Table 3.11: Example of a final cluster list in EXCEL, showing the base forms of the ten most frequently used PVs

Position	***Cluster***	***Absolute frequency***	***Relative frequency****
1	END UP	582	39.06
2	GROW UP	339	22.75
3	CARRY OUT	160	10.74
4	FIND OUT	154	10.34
5	BRING ABOUT	134	8.99
6	GO ON	119	7.99
7	SET OUT	100	6.71
8	STAND OUT	98	6.58
9	GIVE OUT	97	6.51
10	SEND OUT	72	4.83

*Per million words

Normalisation was then applied and reported in the *relative frequency* column. The concept of normalisation, or "frequency per million words", forms an important part of the discussion of the findings, and it is, consequently, suitable to provide clarification of the process at this point. Absolute token frequency (raw count) can only be used when comparison between corpora / subcorpora is not required (McEnery *et al.*, 2006:53). However, in this study, for the sake of the validity of the research, normalisation was necessary because corpora of different sizes needed to be compared. According to McEnery and Wilson (2001:83):

> It should be noted that … if the sample sizes on which a count is based are different, then simple arithmetical frequency counts cannot be compared directly with one

> another: it is necessary in those cases to normalise the data using some indicator of proportion.

They further state that, while there are many ways in which proportion may be indicated, they all depend on "ratio between the size and sample and the number of occurrences of the type under investigation" (McEnery & Wilson, 2001:83). A further factor needs consideration, in that with large corpora this calculation might result in small numbers that are awkward to deal with. Thus, McEnery and Wilson (2001:83) advise multiplying the ratio formula by a large constant. For example, in their research, Biber *et al.* (1999:38) used a million words of text as their basis for normalisation when normalising LSWE frequency counts, a practice that was adhered to here.

Consequently, the following formula was used for normalisation:

x(1 000 000 / y)*

where *x = absolute frequency of the PV* and *y = total number of tokens in the corpus / subcorpus*. Thus, for example, for the PV *end up* with a frequency of 582, and with the token count of the corpus (in this case, the NWU subcorpus) of 14 898 759, the formula would read:

582(1 000 000 / 14 898 759)*

Therefore, an absolute frequency of 582 for the PV *end up* produces a frequency per million words (relative frequency) of 39.06, as shown in Table 3.11 above. In contrast to the presentation of relative frequencies in the pilot study, where relative frequencies were rounded up to the nearest integer, these frequencies were rounded to two decimal places for the main study, as this was felt to provide a more precise comparison of values.

3.7.2.3 *Report on PV distribution within subcorpus*

Once all PVs had been verified, and the absolute and relative frequencies calculated, the PVs were sorted by absolute frequency (highest to lowest). This provided information about the spread of PVs within the subcorpus.

While the standard of 40 or more occurrences per million words had seemed applicable in the pilot study, it was found that PV frequencies in the SAMuLCAT subcorpora displayed notably lower PV use. Another marker was thus needed for PV use comparison among the various subcorpora, in order to facilitate discussion around PV use. The patterns of PV use displayed by the extracted data suggested that PV use might effectively be separated into three frequency groups: PVs with five or more occurrences per million words, PVs with between one and five occurrences per million words, and PVs with one or fewer occurrences per million words. This opened up further interesting comparison possibilities.

The information derived from each subcorpus during the data processing was summarised according to the categories discussed in the previous paragraph, and presented in the format displayed in Table 3.12. The relative frequency of PV tokens per frequency group, as well as the relative frequency of the PV types that make up the PV token frequency at that level, was included in the table. These figures were also given as a percentage of total PV tokens and types. According to Zhou and Wang (2024:4), "researchers often report the percentage of PVs in the total wordcount of the corpus, to showcase the common/uncommon presence of PVs".

Table 3.12: Example of summarised distribution of PVs per frequency group

Frequency group	***Relative frequency* of PV tokens in this group***	***Relative frequency as percentage** of total PV tokens***	***Relative frequency* of PV types represented in this group***	***Relative frequency as percentage** of total PV types***
Five or more occurrences per million words	119.67	41.69%	0.60	2.50%
Between one and five occurrences per million words	101.62	35.40%	3.49	14.50%
One or fewer occurrences per million words	65.78	22.91%	20.00	83.00%
Total	**287.07**	**100%**	**24.10**	**100%**

*Per million words *Of normalised values

The following information can be extrapolated from Table 3.12. While the largest percentage of PV tokens (41.69%) occurred five or more times per million words, this represented only 2.50% of PV types. Therefore, a small percentage of specific PVs occurred reasonably often. On the other hand,

only 22.91% of PVs occurred once or less, although this represented 83% of PV types. Hence, most of the PVs in the subcorpus occurred once only. The conclusion that can be drawn from this is that the students represented in this subcorpus, while familiar with a small number of PVs that they used reasonably often, generally used PVs only once, suggesting an unfamiliarity with general PV use, or a reluctance to incorporate PVs in their academic writing.

3.7.2.4 *Extraction of ten most frequently used PVs*

The sorting of the final PV list not only demonstrated the distribution of PVs, but also provided a list of the most frequently used PVs within the subcorpus. Zhou and Wang (2024:4) state that researchers are not only interested in the total number of PVs in a subcorpus, but also in the PVs that occur most often. The most suitable number of frequently used PVs to extract will depend on the size of the subcorpus (Zhou & Wang, 2024:4). In this case, the ten highest frequency PVs per subcorpus were displayed (in table format).

The most frequently occurring PVs are of particular interest for reasons of comparison. Examining often used PVs across institutions, for instance, makes it possible to assess whether South African students comply with the pattern observed by Chen (2013a:421), that "more high frequency phrasal verbs [identified in a native English corpus (BNC)] are used by most of the learners in a productive test than the less frequent ones". In the case of the present study, this would mean determining whether the South African L2 students represented in the various subcorpora were inclined to use the same high frequency PVs as the L1 students. Furthermore, "divergences such as underuse or overuse between the L2 learners' and English L1 speakers' patterns" may be identified (Zhou & Wang, 2024:4).

Table 3.13 provides an illustration of how information on the ten most frequently used PVs was presented. The absolute and relative frequencies for each PV were included. Comparisons would afterwards be drawn between an institution's L1 and L2 most used PVs, and then across the three institutions. Finally, the PVs most used by L1 and L2 students were compared. The table of the ten

most frequently used PVs was also used to establish which ALTs might have been used by students in preference to PVs. The selection of ALTs is discussed in the next section (§3.8.2.5).

Table 3.13: Example of list of ten most frequently used PVs

PV	***Absolute frequency (includes all versions of the base form of the PV)***	***Relative frequency****
END UP	582	39.06
GROW UP	339	22.75
CARRY OUT	160	10.74
FIND OUT	154	10.34
BRING ABOUT	134	8.99
GO ON	119	7.99
SET OUT	100	6.71
STAND OUT	98	6.58
GIVE OUT	97	6.51
SEND OUT	72	4.83

*Per million words

3.7.2.5 *Comparison of ALTs*

Students' use of PVs next had to be compared to their use of ALTs, so as to ascertain to what extent they were avoiding the use of PVs in favour of ALTs. According to Bybee (2006:720), the more frequently a construction is used, the more it is "processed as a unit rather than through its individual parts" and "grows autonomous from the construction that originally gave rise to it". Therefore, high frequency of use creates a familiarity with an expression in the mind of the user that surpasses the sum of its parts. It is thus probable that L2 students would be less likely to avoid the PVs with which they are most familiar. However, because of time constraints it was not possible to check for ALTs for the vast number of infrequently used PVs in a particular subcorpus. Therefore, it was decided to investigate the use of ALTs for the ten most frequently used PVs, as it was felt that reasonable inferences could be made about the subcorpus based on this information.

As no reference work exists that provides suitable ALTs for PVs, several methods were used to compile a useable ALT reference list. Firstly, the two PV dictionaries previously used to identify PVs were consulted again, as they both contained a list of PVs and ALTs as appendices. These lists were useful

but not exhaustive, which necessitated the use of other supplementary sources. The *Collins COBUILD Phrasal Verbs Dictionary* (2013) also occasionally included an ALT in the PV definitions. To this was added the ALT options found in the research articles that had been consulted previously (Dagut & Laufer, 1985:74; Haider *et al.*, 2020:1189; Laufer & Eliasson, 1993:38). Further, if a PV was encountered for which there was no ALT as yet, a possible ALT was sought from the online dictionaries, and the PV and ALT pair were then added to the reference list.

Creating a list of synonyms is complicated. As was apparent from the discussion on the many sense meanings that a PV might have, PVs cannot simply be replaced by an ALT without taking the sense in which the PV is being used into consideration. Therefore, to check the frequency of use of the ALTs for the ten most frequently used PVs for each of the subcorpora, each concordance line ideally had to be inspected to discover the correct sense meaning of the PVs, as determining the key meanings of the relevant PVs "requires reading its tokens in context" (Liu & Myers, 2020:410). However, this proved to be an almost insurmountable task, given the, at times, high PV frequencies. The researcher, therefore, attempted to obtain a general idea of the various ways in which a PV had been used by scanning the context of only a selected number of concordance lines per PV. Normally, the lines were scanned until approximately three or four different meanings had been encountered for that PV. If the PV displayed great diversity, more lines were scanned.

Thereafter, the most appropriate ALT, or ALTs, was selected from the reference list, or the online dictionaries once again consulted to see if a better option could be found, if necessary. For many PVs, the most appropriate synonym would be the more formal ALTs that many seem to prefer to use in academic writing. On the other hand, some PVs simply do not have a suitable synonym. As Chen (2013a:436) rightly comments, "not every phrasal verb has a single-word equivalent and vice versa".

Once the appropriate ALTs had been selected, the Wordsmith Tools Concord option was used to determine their frequency of use. For each PV, several ALTs might be tried, yet only the three most prolific options reported on. For example, for the PV *end up* only one ALT, namely *arrive*, could be

found. Yet even this ALT proved to be somewhat unsuitable, as it (nor its various permutations) often did not appear in the subcorpora. On the other hand, in some cases, several synonyms might be considered. For instance, *project*, *protrude*, *surpass*, and *excel* were thought to be suitable ALTs for the PV *stand out*. However, only the three synonyms with the highest number of occurrences would then be included in the table.

The PV *grow up* presented an interesting problem. For this PV, two ALTs were available, namely *mature* and *develop*. When checked against the use of *grow up* in the NWU L2 concordance list, it appeared that both options were possible ALTs. For example, in the sentence '*…together to influence how he / she will* ***grow up*** *and develop…*', the PV *grow up* could have been substituted by the verb *mature*. Likewise, in the sentence '*…to learn new things. Students can* ***grow up*** *as good communicators in addition to…*', the PV *grow up* could have been substituted by the verb *develop*. However, such examples of appropriate uses of the ALTs were limited. Instead, it was generally found that the PV *grow up* could not reasonably be replaced by either of the two options, as illustrated by the sentence '*…to kill themselves and that will make them* ***grow up*** *with anger - and they will be…*'.

In many cases, a substantial amount of refining of the resultant concordance lines was needed. One instance of this might be that all non-verb variations of the ALT would need to be removed. For example, the word *mature* often featured as an adjective (*mature* people, *mature* adults, and *mature* users). Another example where refining might be necessary would be in the case of the ALT *develop*, which produced a concordance list of 3 334 lines, of which only 22 were viable alternative options. (This means that the ALT could only be effectively replaced by the PV *grow up* in 22 cases.) This result confirmed the necessity of working through the ALT concordance lines. The meaning of the concordance line was also on occasion unclear, as in '*… provided with internet use hence it presents bullies with a method to* ***undertake*** *in domineering behaviours on a larger spectrum without…*'. If, in such cases, the intended meaning could not be determined, the line was dropped. While due care was

taken, this process, being manual, cannot be claimed to be without error. Yet it is hoped that such errors were limited in number and did not substantially impact the results.

While the scanning of the ALT concordance lines was manageable for most of the subcorpora, this was not always the case. For the larger subcorpora, such as the NWU L2 subcorpus, up to 5 678 lines were produced per ALT. Because of the time it would take to validate each line individually, the process was expedited by looking for patterns that would identify as valid more effectively. For instance, the ALT *develop* was found to only be acceptable as an ALT for the PV *grow up* when referring to *children* or *learners*, as in '*…As explained by Piaget (1954) children, while* ***developing****, are influenced by external factors and experiences*'. Therefore, these two words (*children* or *learners*) were searched for and the resulting lines checked for validity. While it is possible that, by using this method, some lines might have been missed, the overall impact would most likely not have been measurably affected.

Liu and Myers (2020:411), likewise, found that, in the case of a large corpus, "it was not feasible to read all [the] token concordance lines as the task would require an enormous amount of time and effort that few researchers could afford". Thus, they employed a procedure similar to that of Garnier and Schmitt (2015) by reading, for example, 200 randomly selected concordance lines and comparing those results to another 200 randomly selected concordance lines. A high level of similarity between these two sets of data suggested that the selected concordance lines "were representative of all the tokens" (Liu & Myers, 2020:411-412). By this means, it was possible to achieve an acceptable level of validation within an acceptable timeframe. Establishing patterns of use by which a large corpus might be effectively navigated is, therefore, not without precedent.

The valid occurrences of the ALTs were totalled and presented in a table. For illustration purposes, Table 3.14 is a shortened form of what such a table might look like. The table gives only one ALT for the PV *end up*, namely *arrive*, and also demonstrates the likelihood of a low frequency for an ALT that is, in fact, not very suitable. Likewise, the PV *grow up* has only two ALTs (*develop* and *mature*) for which there are again low frequencies. Both these PVs are highlighted in light grey as they have higher

frequencies than their ALTs. On the other hand, the PV *carry out* has three possible ALTs, *perform*, *execute* and *undertake*. Of these three, *perform* has a substantially higher frequency than its PV counterpart, and is highlighted in light grey.

Table 3.14: Example of ALTs for ten most frequently used PVs

PV	Absolute frequency	Relative frequency*	ALT	Absolute frequency	Relative frequency*
END UP	582	39.06	Arrive	0	0.00
GROW UP	339	22.75	Develop	17	1.14
			Mature	5	0.34
CARRY OUT	160	10.74	Perform	250	16.78
			Execute	91	6.11
			Undertake	50	3.36

*Per million words

3.7.2.6 *Error processing*

During the process of data extraction and analysis, PVs that were considered to have been incorrectly used were flagged. The concept of errors in language use can be seen as contentious. For example, Boughey (1998:171) maintains that errors should not simply be seen as a "lack of linguistic awareness" but that there should also be an attempt "to look to other reasons for those errors being made". Furthermore, she quite correctly states that:

> In the process of writing, even educated native speakers of English produce many sentences which are grammatically incorrect. Those errors have not been made because of a lack of knowledge of English grammar but rather because writers have not yet discovered what it is that they want or mean to say.

Why, then, report errors at all? One reason is that, though proficiency is suggested by the presence of PVs in student writing, this can only be confirmed if the PVs have been used correctly (Chen, 2013b:91-92). Moreover, by indicating specific areas of concern, PV error reporting can prove useful to educators. This is in contrast to the view that language errors simply need to be corrected. Rather, they are now seen as "important indicators to facilitate the process of learning a language", and, therefore, investigating errors has become indispensable to language learning (Özkayran & Yilmaz,

2020:50). This is in line with the generally acknowledged need for the improvement of English language proficiency at South African universities (Van Rooy & Coetzee-Van Rooy, 2015:32). A further argument is that indicating anomalies in L2 student writing is not necessarily aimed at pointing out the lack of proficiency of L2 students, especially if L1 errors are also reported (as was done in this research). In this regard, it should be pointed out that differences (as in the differences between L1 and L2 student PV use) are not deficiencies.

The context within which PVs are reported as incorrect should be clarified at the outset. The rules that were applied were based on the various hardcopy and online dictionaries that were referenced during the qualitative analysis process. The definition and legitimate use of each PV type was checked against these reference works. Establishing legitimacy at times also included referring to the use of the PV in context (reading the relevant concordance line). In the case of the use of an invalid PV, informal sources were also referenced (see discussion on *unconfirmed PV* below).

Many different kinds of PV errors have been identified in PV research, some of which have an influence on the quantity of PVs produced, such as avoidance and "style deficiency", and others which affect the quality of the PVs produced, such as semantic and syntactic errors (Chen, 2013b:91). Avoidance (as discussed in §2.8) points to a reluctance by L2 students to use a construction with which they are unfamiliar, whereas "style deficiency" might point to an overuse of PVs, especially where a more formal style is required. The errors discussed below fall within the semantic and syntactic categories. Examples of the various anomalies encountered are given in Table 3.15.

Table 3.15: Kinds of PV anomalies found in concordance lines

PV	***Concordance line***	***Error***
Voice up	"…ways. Firstly ,[1] victim student should **voice up** and inform school staff member and parents…"	Non-existent PV
Drive up	"…This type **drives up** the cost for government…"	Unconfirmed PV
Cast down	"…be and vice versa. Such beliefs have been **cast down**…"	Incorrect PV use
Give in	"…countries can also follow the same initiation to **give in** a helping hand. Word…"	Redundant particle
Hang up	"…where children no longer feel comfortable staying or **hanging up** with other children because they are al…"	Incorrect particle
Write up	"…the love of creative learning in this **write - up** we will be looking at the…"	Noun, not PV
Break down	"…explained during the course of this essay to give you the **broken down** understanding…"	Adjective, not PV

1 Please note that student errors have been copied as found, and have therefore not been corrected.

As discussed previously, the method of error extraction is a by-product of PV validation, in that the errors observed while checking concordance lines for PV use were noted and reported on. Scrutiny of the concordance lines also provided the opportunity to discard words that had erroneously been tagged as *verb + particle* combinations, while being neither PVs nor PV errors. Two examples of such errors are *broken down,* which was found to be an adjective rather than a PV ('*…the* ***broken-down*** *understanding…*'), and *write up*, which was found to be a noun rather than a PV ('…*this* ***write-up***…'). Instances such as these were not reported as errors in the research, as they were the result of incorrect tagging rather than incorrect PV use.

PV errors were classified as follows. Firstly, errors were identified when a *verb + particle* combination seemingly adhered to the definition of a PV, but could not be verified in any recognized reference work, as in the example of *voice up*. This kind of error was categorised as a n*on-existent PV*. Nevertheless, in some instances, such PVs were found to be referenced informally online (as in the example *drive up*), which suggested that they might in future become formally incorporated into the language. This was then indicated as a separate kind of error from the *non-existent PV*, namely that of u*nconfirmed PV*. On the other hand, a legitimate PV might have been used incorrectly, which was labelled as *incorrect PV use*. For example, the PV *cast down* is a legitimate PV, but has been used

incorrectly in the phrase *'*Such beliefs have been **cast down**…*'. (Indeed, it is not clear what the student meant to say here - perhaps, *rejected*?).

These three kinds of errors might be considered developmental errors, as they appear to be indicative of an attempt to apply language rules already learnt to create new PVs, or to apply known PVs to new situations. As reported in Chapter 2, the PV is a highly productive grammatical form, with L1 speakers coining new PVs on a regular basis. Therefore, it does not seem unreasonable that L2 students should be tempted to do the same.

The next two kinds of error can be discussed together, as they both relate to the use of the particle. In the case of the error labelled r*edundant particle*, an unnecessary particle was found to have been added (*'*…to **give in** a helping hand…*' rather than '*…to **give** a helping hand…*'). Thus, a verb that adequately performed its role had been needlessly turned into a PV by the addition of a particle. On the other hand, in the example *'*…**hanging up** with other children…*', the particle is clearly incorrect because the phrase should read '*…**hanging out** with other children…*'. This kind of error was labelled i*ncorrect particle*.

The first three PV error examples in the table can be classed as semantic, as they pertain to meaning. The next two PV error examples are syntactic in nature, as they indicate that a grammatical structure (the PV) has been used incorrectly. These two categories of error are relevant to the research aims of this study.

Table 3.16 illustrates how error reporting was presented. As discussed above, there were five kinds of errors that could be reported on for each of the subcorpora, although not all of the different kinds of errors were present in each subcorpus. Because it was necessary to distinguish between the kind or PV errors that students made and the number of times a particular kind of error was made, it was decided to use the same terminology as that used for distinguishing between the base form of PVs, and the number of times the base form occurred, namely *types* and *tokens*. This was deemed suitable

as it adheres to *type* referring to "a representative instance", and token referring to "each individual occurrence of a particular type", as defined by Weisser (2016:149). Errors were, therefore, totalled in two ways (*Error type frequency* and *Error token frequency*), and normalised totals were also provided (*Relative error type frequency* and *Relative error token frequency*). The difference between the two totals is discussed in more detail in the next paragraph.

Table 3.16: Example of error reporting

Type of PV error	***Error type frequency***	***Relative error type frequency****	***Error token frequency***	***Relative error token frequency****
Redundant particle	54	3.63	108	7.25
Incorrect PV use	46	3.09	82	5.50
Non-existent PV	16	1.07	20	1.34
Incorrect particle	9	0.60	11	0.74
Unconfirmed PV	8	0.54	17	1.14
Total	**133**	**8.93**	**238**	**15.97**

*Per million words

Table 3.17 illustrates the difference between *error type frequency* and *error token frequency*. Firstly, the total for *error type frequency* refers to the number of times this kind of error was identified. For example, the table shows four incidences of an *unconfirmed PV. Y*et the *error token frequency* column indicates that some specific examples of this type of error occurred more than once. For instance, *drive up* appeared nine times and *lock down* appeared twice. This discrepancy between the *error type frequency* and *error token frequency* indicates that some specific error type examples (such as, in this case, *drive up*) might have become reasonably well established in the minds of the students.

Table 3.17: Example of EXCEL sheet showing error reporting

Cluster	***Error***	***Error type frequency***	***Error token frequency***
MAKE BACK	**Unconfirmed PV**	1	1
DRIVE UP	**Unconfirmed PV**	1	9
LOCK DOWN	**Unconfirmed PV**	1	2
SEND THROUGH	**Unconfirmed PV**	1	1

In Chapter 4, error reporting per subcorpus is summarised per error type and presented in table format, of which Table 3.18 is an example. Besides totals for error type frequency and error token

frequency being provided, these totals are also normalised per million words for purposes of valid comparison between L1 and L2 groups within subcorpora, and between institutions.

Table 3.18 Example of error reporting per subcorpus

Type of PV error	*Error type frequency*	*Relative error type frequency**	*Error token frequency*	*Relative error token frequency**
Redundant particle	21	4.45	34	7.21
Non-existent PV	15	3.18	15	3.18
Incorrect PV use	14	2.97	23	4.88
Incorrect particle	11	2.33	15	3.18
Unconfirmed PV	3	0.64	5	1.06
Total	**64**	**13.58**	**92**	**19.52**

*Per million words

Finally, it should be noted that error frequencies were not included in the PV token frequency, to ensure valid PV frequency reporting.

3.7.3 Analysis process

The data processing described in the previous sections were applied as follows. The PV use of the L1 and L2 subcorpora of each institution was investigated separately, after which they were compared to each other. This provided an overview of the differences in PV use between L1 and L2 students within the institution, which could be reported on according to the parameters previously described: PV type and token frequencies, the distribution of PVs within the subcorpus, the highest frequency PVs, the use of ALTs to PVs, and PV errors. The process was repeated for each of the three institutions, in the sequence NWU, UP and WITS.

Once the L1 and L2 PV data for each of the three universities had been extracted, the information was collated in order to compare PV use across institutions. Again, the same parameters were used for reporting. Relative frequencies were used to maintain validity of measurement, although a comparison of overall token count was included to indicate the differences in subcorpora sizes. According to Brezina (2018:270), it is necessary to be aware of the sizes of the subcorpora that are to be compared as far as token count is concerned, but that the focus should be on the “occurrences of

the target linguistic feature" within the subcorpora. Subcorpus token count was, therefore, mentioned in each case, but the main aim was to report on PV use, using relative values because of the varying sizes of the subcorpora.

In this regard, it should be noted that all possible PV combinations were extracted from the relevant corpora, after which the combinations were verified so that only valid PVs were retained. The total number of occurrences (tokens) of each valid PV (types) was then noted. (See §3.5 for clarification of the use of the terms *type* and *token* in this study.) Normalisation was done on the extracted PVs (types and tokens). Therefore, the normalisation was done with exact figures, and was not used to project the number of further PV types that might be encountered. The normalisation process, in this case, presents an average of PV types per million words of a corpus, which can be checked against the total PV types for the corpus.

Comparing PV use across universities made it possible to determine whether patterns of L1 and L2 PV use were similar or different at the different universities, as well as whether, overall, patterns of PV use were similar or different at different universities. The comparison included an investigation into the manner in which they differed, if at all. Whether the PV use at an institution aligned with its ranking, as had been surmised, was also investigated.

The final step was to compare overall L1 and L2 PV use. Indeed, this formed the main focus of the research. The relative frequencies of the information gathered at institutional level were collated for use in comparing L1 and L2 PV use, employing the same parameters as before. This made it possible to address the research questions that centred around whether L1 or L2 students used more PVs, whether L1 or L2 students used more ALTs, and whether L1 or L2 students were more likely to adhere to syntactic and semantic norms in their use of PVs.

Descriptive statistics were used in this research, being the most practical for describing the results. This is in keeping with the statement by McEnery *et al.* (2006:52) that, "most users of corpus data will

not be capable of generating sophisticated statistical claims nor would they wish to do so". Descriptive statistics are used to "reveal the main tendencies in the dataset" by means of overviews and simple graphs, in order to produce "a reliable description of the data" (Brezina, 2018:259).

3.8 Conclusion

In this chapter, the problem statement and research aims of the present research were clarified, as well as the research design in which the research would be embedded. A discussion of the pilot study gave an overview of the format in which the research was to be presented. Thereafter, the corpora to be used in the present study were described. The methods by which the data would be collected and processed were then explained.

In the next chapter, the analysis of the data is presented. Each institution (NWU, UP, and WITS) will first be analysed individually by means of a comparison of L1 and L2 PV data, after which PV use across the institutions will be compared, followed by a comparison of overall L1 and L2 PV use.

Chapter 4: Results and analysis

4.1 Introduction

The main aim of this research study was to investigate PV use by L1 and L2 students. This was made possible by the fact that the SAMuLCAT and WITS corpora included the writings of L1 and L2 students. Furthermore, it was also possible to compare PV use across institutions, as the student data from different universities were represented in the corpora. The procedures used to extract the data were described in detail in the methodology section. In brief, Wordsmith Tools was used to generate the required frequency lists per institution in order to investigate patterns of use. Both quantitative and qualitative analyses were used. The generation of frequency lists is a quantitative process, while the manual verification, required to eliminate from the frequency lists those multi-word combinations that are not PVs, is qualitative in nature.

In this chapter, the extracted data are reported on and analysed according to the research questions discussed in §3.3.1. The primary task is to identify variance in the patterns of use (Botha, 2012:72) among the subject groups, and thereafter to investigate what the patterns reveal about PV use by South African L2 students in comparison to L1 students, whether L1 and L2 students prefer the use of PVs or the use of one-word lexical items that are alternatives to PVs (ALTS), and whether L2 students' use conforms to standard grammar rules.

Addressing the research questions pertaining to overall L1 and L2 PV use is accomplished by means of the following process. Firstly, L1 and L2 PV use at each institution (NWU, UP and WITS) is investigated separately according to certain parameters (PV type and token frequency, PV distribution, most frequently used PVs, possible preference for ALTs, and PV error frequency), after which L1 and L2 use within the institution is compared. While not addressing particular research questions, this approach makes it possible to observe preliminary patterns of PV use within the institutions. Thereafter, L1 and L2 PV use across institutions is compared and reported on, facilitating

a response to the research questions based on differences in PV use among L1 and L2 students from diverse universities. Finally, the L1 and L2 PV data that were collected during the process are collated and reported on.

As discussed in §3.5, the terms PV *type* and *token* will be used according to Weisser's (2016:149) definitions, with *type* referring to "a representative instance/word form in a frequency list", and *token* referring to "each individual occurrence of a particular type".

4.2 Analysis of L1 and L2 PV use within three South African tertiary institutions

As explained in the introduction to this chapter, PV use at the three institutions (North-West University, University of Pretoria, and University of the Witwatersrand) will be discussed individually, and the L1 and L2 PV use per institution will be examined and reported on separately. This allows for the observation of particular patterns of PV use by different groups of students. Thereafter, a comparison will be drawn between the L1 and L2 groups to highlight differences in use, if any. The L2 group will be examined first, as L2 PV use has formed the focus of most previous research into PV use. In each case, data will be measured using the same parameters. These are PV type and token frequencies, PV distribution within the subcorpus, the ten most frequently used PVs, ALT frequencies, and PV error frequencies. Figure 4.1 provides an outline of the structure of this section.

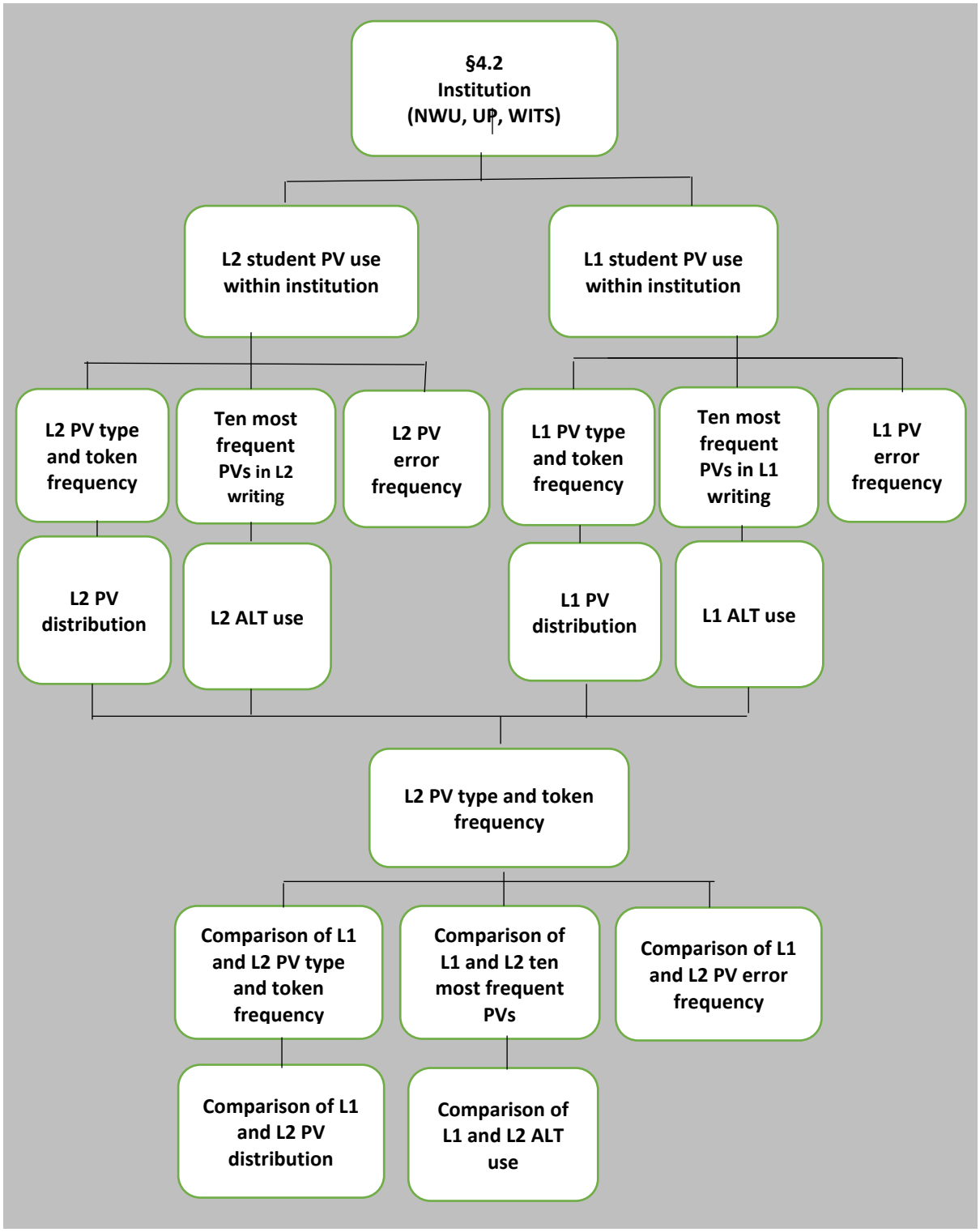

Figure 4.1 Structure of subcorpora analysis

Note that, in the discussion of the texts, tagging information was removed from concordance line examples to facilitate reading. For instance, the line '*…happen . A hacker known as Paladin ,* type='VBD'>*was able to* <w type='VB'>*shut* <w type='RP'>*down several government*…' will be changed to '*…happen . A hacker known as Paladin , was able to shut down several government*…'. Errors in

student writing have not been corrected (such as incorrect spacing around punctuation, as evident in this example), to retain the authenticity of the student writing.

4.2.1 North-West University (NWU)

The first institution to be considered is the largest subcorpus in the SAMuLCAT corpus as well as the largest of all the subcorpora used in this research, with 14 898 759 L2 tokens, 2 030 568 L1 tokens, and 16 929 327 tokens in total, as shown in Table 4.1 below. The L1 tokens make up 12% of the subcorpus, while the L2 tokens make up 88%.

Table 4.1: NWU L1 and L2 subcorpora

Subcorpus	*Subcorpus size (tokens)*	*Percentage of total subcorpus size*
L1	2 030 568	12%
L2	14 898 759	88%
Total	**16 929 327**	**100%**

4.2.1.1 *NWU L2 student PV use*

The SAMuLCAT NWU L2 subcorpus produced 4 277 valid PV tokens, including all versions of the base form of the PV (see Table 4.2 below). These 4 277 PVs, with a relative frequency of 287.07 (normalisation applied, as discussed in §3.8.2.2), were made up of 359 PV types, with a relative frequency of 24.10. As discussed in §3.5, *absolute frequency* is the total number of times a particular word appears in a corpus, while *relative frequency* refers to the normalisation of the absolute frequency.

Table 4.2: Total PV frequency in NWU L2 student writing

Description	*Absolute frequency*	*Relative frequency**
PV types	359	24.10
PV tokens	4 277	287.07

*Per million words

4.2.1.1.1 Distribution of PVs in NWU L2 subcorpus

A grouped frequency table is used to organise the data into categories, or classes, that make the data more manageable. In this study, therefore, the data are presented in a grouped frequency table,

illustrated in Table 4.3 below, which shows the number of PVs per class interval, or frequency group, in the NWU L2 subcorpus. The frequency groups used here are: five or more occurrences per million words, between one and five occurrences per million words, and one or fewer occurrences per million words (see §3.8.2.3).

Table 4.3: Distribution of PVs per frequency group in NWU L2 student writing

Frequency group	***Relative frequency* of PV tokens in this group***	***Relative frequency as percentage** of total PV tokens***	***Relative frequency* of PV types represented in this group***	***Relative frequency as percentage** of total PV types***
Five or more occurrences per million words	119.67	41.69%	0.60	2.50%
Between one and five occurrences per million words	101.62	35.40%	3.49	14.50%
One or fewer occurrences per million words	65.78	22.91%	20.00	83.00%
Total	**287.07**	**100%**	**24.10**	**100%**

*Per million words **Of normalised values

The data show that 41.69% of PV tokens (*PV tokens within frequency group / total PV tokens * 100*) occurred five times or more times per million words, which is the highest occurrence percentage of all the frequency groups. However, only a small percentage of PV types (2.50%) is represented in this category. The percentage of PV types was calculated using the formula: *PV types within frequency group / total PV types * 100*. The comparatively small percentage (2.50%) suggests that a small number of PVs were used reasonably consistently.

In the next category (*between one and five occurrences per million words*), 35.40% of the total PVs were used. Again, only a comparatively small percentage of PV types (14.50%) is represented, yet this percentage is approximately six times that of the percentage of PV types of the previous category. The smallest percentage of PV tokens (22.91%) falls within the *one or fewer occurrences per million words* category. Nevertheless, this represents 83% of PV types used. This result suggests that, in this subcorpus, PVs are generally used once only, and are, therefore, not a regular feature in NWU L2 student writing. Thus, it is likely that the use of ALTs will be found to be more prevalent than PVs in the NWU L2 subcorpus.

4.2.1.1.2 Ten most frequently used PVs in NWU L2 subcorpus

The ten PVs with the highest frequency (representing all versions of the base form of the PV) in the subcorpus are presented in Table 4.4, along with their relative frequencies. According to Liu and Myers (2020:405), creating lists of the most used PVs is useful and necessary. This is because the number of PVs in the English language is so large (more than 10 000) that concentrating on the most used PVs is more practical than trying to incorporate all possible PVs in one's teaching.

Table 4.4: Ten most frequently used PVs in NWU L2 student writing

PV	***Absolute frequency***	***Relative frequency****
END UP	582	39.06
GROW UP	339	22.75
CARRY OUT	160	10.74
FIND OUT	154	10.34
BRING ABOUT	134	8.99
GO ON	119	7.99
SET OUT	100	6.71
STAND OUT	98	6.58
GIVE OUT	97	6.51
SEND OUT	72	4.83

*Per million words

The relative frequencies of these ten PVs show a sharp drop in use, from 39.06 to 4.83, a difference of 34.23. Such a "rapidly diminishing word frequency" is called Zipf's law (Brezina, 2018:44). Brezina's (2018:44) informal explanation of this law is:

> when we start with the most frequent item in the wordlist (regardless of the size of the corpus), the second most frequent item will have only half of the frequency of the first item. The third most common word will have one-third of the frequency of the first item, and so on.

However, in reality, although this general pattern will be observed, the rate at which the frequencies drop cannot be predicted with complete accuracy (Brezina, 2018:46). The point of Zipf's law is rather to be aware that the evidence available about word behaviour in a corpus suggests that frequencies

diminish at a rapid rate, which means that it is important to "critically evaluate the amount of evidence we have for our claims" (Brezina, 2018:46).

Though not errors, it might be interesting to note the use of some slang PVs that were used in this subcorpus, such as *rev up* and *churn out*.

4.2.1.1.3 Use of ALTs in the NWU L2 subcorpus

Using the ALT list that had been created previously, the most appropriate ALTs were selected for frequently used PVs in the NWU L2 subcorpus. Wordsmith Tools was then used to find the occurrences of the ALTs in the subcorpus. The results are given in Table 4.5.

Table 4.5: ALTs for the ten most frequently used PVs in NWU L2 student writing

PV	Absolute frequency	Relative frequency*	ALT	Absolute frequency	Relative frequency*
END UP	582	39.06	Arrive	0	0.00
GROW UP	339	22.75	Develop	17	1.14
			Mature	5	0.34
CARRY OUT	160	10.74	Perform	250	16.78
			Execute	91	6.11
			Undertake	50	3.36
FIND OUT	154	10.34	Learn	216	14.50
			Discover	166	11.14
			Detect	6	0.40
BRING ABOUT	134	8.99	Cause	1500	100.68
			Generate	192	12.89
			Produce	189	12.69
GO ON	119	7.99	Continue	181	12.15
			Proceed	34	2.28
SET OUT	100	6.71	Explain	956	64.17
			Outline	172	11.54
			Arrange	21	1.41
STAND OUT	98	6.58	Excel	28	1.88
			Protrude	Not relevant**	0.00
			Surpass	Not relevant	0.00
GIVE OUT	97	6.51	Distribute	108	7.25
			Issue	28	1.88
			Provide	Not relevant	0.00
SEND OUT	72	4.83	Distribute	75	5.03
			Emit	45	3.02
			Issue	18	1.21
Total	**1855**	**124.51**	***Total***	**4348**	**291.84**

*Per million words **This means that the synonym proved to be an inadequate alternative to the PV.

Table 4.5 shows that only three PVs were preferred to ALTs (indicated by light grey shading). These PVs were *end up*, *grow up*, and *stand out*. In all three cases, no suitable alternative could be found, which may have influenced the students' preference for the PVs. It should be noted that, in some cases, not all of the ALT options considered were preferred to PVs. For example, in the case of *carry out*, only the ALT *perform* was preferred to the PV. Overall, ALTs (with a relative total of 291.84) were

preferred to PVs (with a relative total of 124.51), by a margin of 134.39 per cent. (For clarification on the ALT selection process, see §3.8.2.5.)

4.2.1.1.4 Errors in NWU L2 PV use

Table 4.6 shows the errors in PV use that were found in NWU L2 student writing. The error classification devised for this study was applied as rigorously as possible in order to ensure consistency (see §3.8.2.6). Nevertheless, the researcher acknowledges that arguments could be made for some of the incorrectly used PVs to be reclassified, as there is a degree of overlap between a few of the categories. For instance, the use of an incorrect particle, as illustrated by the example *use out*, could also have been seen as a non-existent PV. As in the case of PV types and tokens, *error type* and *token* are used here as described in the methodology chapter (§3.8.2.6), with *type*, in this case, referring to "a representative instance", and *token* referring to "each individual occurrence of a particular type" (Weisser, 2016:149). (See the discussion of the error *issue out* below for an example of the difference between error type and error token.)

Table 4.6: Errors in PV use in NWU L2 student writing

Type of PV error	***Error type frequency***	***Relative error type frequency****	***Error token frequency***	***Relative error token frequency****
Redundant particle	54	3.63	108	7.25
Incorrect PV use	46	3.09	82	5.50
Non-existent PV	16	1.07	20	1.34
Incorrect particle	9	0.60	11	0.74
Unconfirmed PV	8	0.54	17	1.14
Total	**133**	**8.93**	**238**	**15.97**

*Per million words

Both absolute and relative values are given for error type and error token, namely *Error type frequency* and *Error token frequency.* The *Error type frequency* column gives the total number of times a *type* of error occurred. For example, there were 54 occurrences of the error type identified as *redundant particle*. In some instances, a particular verb and redundant particle combination occurred more than once, as in the case of *issue out*, which appeared 12 times. Thus, the difference in the two totals (*Error*

type frequency and *Error token frequency*) suggests that some particular examples of error types were used more than once by students. The relative frequencies of these two columns are of most interest to us, being the normalised totals with which subcorpora can legitimately be compared. Therefore, reference will primarily be made to the *Relative error type frequency* column, as well as the *Relative error token frequency* column.

The error type that was most prevalent in this subcorpus was the addition of an unnecessary particle. As previously discussed, there were 54 of these types of errors (with a relative frequency of 3.63), and 108 error tokens (or 7.25 occurrences per million words), which means that particular instances of this error appeared more than once. For instance, the incorrect PV *issue out* (as used in the sentence *'…*to implement this method the government must* ***issue out*** *the mandate to the police officials*…') is one of the 54 examples of a particle being unnecessarily added to a verb. This particular version of the error appears 12 times in the subcorpus, adding to the 108 error token frequency. As discussed in §3.8.2.6, this phenomenon might indicate that the particular PV error has become entrenched in student writing.

The next error type that occurred with high frequency was the incorrect use of an existing PV. An example of this kind of error was the PV *turn out*, which occurred three times. While this is a valid PV, its use in the following sentence is not: *'…*to share with other learners and that could* ***turn out*** *to learners misbehaving during the lesson*…'. (The intention might have been to say, '…*and that could* ***result in*** *learners misbehaving*…'). The relative frequency of this type of error was 3.09, and the token frequency was 5.50. Again, the discrepancy between the frequencies is the result of particular instances of the error type appearing more than once, such as, among others, *turn out* (three times), *blow out* (four times), *take in* (five times), and *give out* (six times), all of which, while appearing to be valid PVs, were used incorrectly on closer inspection.

The next two types of PV errors that were found, namely the use of a non-existent PV (such as *attain out*), and the pairing of a verb proper with an incorrect particle (such as *use out*, instead of *use up*)

appeared to a far lesser extent, with relative error type frequencies of 1.07 and 0.60, and with relative error token frequencies of 1.34 and 0.74, respectively.

The final type of error, the unconfirmed PV, with examples such as *make back, lock down, drive up, isolate from*, and *grow back*, had a relative error type frequency of only 0.54, in contrast to a relative error token frequency of 1.14. This is as a result of the PV *drive up* (as in the sentence '…*marketers who develop marketing strategies to* ***drive up*** *sales or services*…') being used nine times, which suggests a general use of a PV that cannot be authenticated as it has not yet been recorded in the dictionaries referred to (See §3.8.2.6). In total, the relative error type frequency is 8.93, and the relative error token frequency is 15.97.

4.2.1.2 *NWU L1 student PV use*

The SAMuLCAT NWU L1 subcorpus produced 546 valid PV tokens (with a relative frequency of 268.89), representing various versions of the base form of the PV, as shown in Table 4.7 below. These 546 PVs are made up of 142 PV types, with a relative frequency of 69.93.

Table 4.7: Total PV frequency in NWU L1 student writing

Description	***Absolute frequency***	***Relative frequency***
PV types	142	69.93
PV tokens	546	268.89

4.2.1.2.1 Distribution of PVs in the NWU L1 subcorpus

As can be seen from Table 4.8, the largest percentage of PV tokens (48.35%, with a relative frequency of 130.01) occurred five times or more. This represents 9.85% PV types (with a relative frequency of 6.89). There was a reasonably high use of a small percentage of PV types. The smallest percentage of PV tokens (20.88%) occurred once or less. However, this represents the largest percentage of PV types (63.39%). Thus, while a small percentage of PVs are used reasonably frequently, most PVs are used once only. Therefore, as in the case of the L2 student corpus, the ALT frequency can be expected to be higher than that of the PV frequency.

Table 4.8: Distribution of PVs per frequency group in NWU L1 student writing

Frequency group	***Relative frequency* of PV tokens in this group***	***Relative frequency as percentage** of total PV tokens***	***Relative frequency* of PV types represented in this group***	***Relative frequency as percentage** of total PV types***
Five or more occurrences per million words	130.01	48.35%	6.89	9.85%
Between one and five occurrences per million words	82.74	30.77%	18.71	26.76%
One or fewer occurrences per million words	56.14	20.88%	44.32	63.39%
Total	**268.89**	**100%**	**69.93**	**100%**

*Per million words *Of normalised values

4.2.1.2.2 Ten most frequently used PVs in the NWU L1 subcorpus

Table 4.9 shows the ten PVs with the highest absolute frequencies, as well as their relative frequencies. It is interesting to note that Zipf's law is not evident here, the data values being more closely distributed than in the NWU L2 subcorpus. The difference in data dispersion patterns between the L1 and L2 subcorpora will be discussed in more detail in §4.2.1.3.2. (The term *dispersion* is used as the "spread of values of a variable in a dataset" (Brezina, 2018:11)).

Table 4.9: Ten most frequently used PVs in NWU L1 student writing

PV	***Absolute frequency***	***Relative frequency****
END UP	29	14.28
CARRY OUT	28	13.79
BRING ABOUT	25	12.31
COME UP	24	11.82
MAKE UP	24	11.82
GROW UP	22	10.83
GO ON	18	8.86
SET OUT	18	8.86
OPEN UP	15	7.39
TAKE ON	14	7.00

*Per million words

4.2.1.2.3 Use of ALTs in NWU L1 subcorpus

Table 4.10 gives the ALTs for the ten most frequently used PVs in the NWU L1 subcorpus. In five of the ten cases, the ALTs were preferred to the PVs (*carry out*, *bring about*, *come up*, *make up*, and *set out*)

(indicated in the table by light grey shading). Reasons for the choice of PV over ALT vary, as will become clear in the following discussion.

Table 4.10: ALTs for the ten most frequently used PVs in NWU L1 student writing

PV	***Absolute frequency***	***Relative frequency****	**ALT**	***Absolute frequency***	***Relative frequency****
END UP	29	14.28	Arrive	0	0
CARRY OUT	28	13.79	Conduct	61	30.04
			Perform	21	10.34
			Execute	13	6.40
BRING ABOUT	25	12.31	Cause	173	85.20
			Produce	74	36.44
			Generate	30	14.77
COME UP	24	11.82	Occur	178	87.66
			Arise	34	16.74
			Emerge	12	5.91
MAKE UP	24	11.82	Create	404	198.96
			Constitute	37	18.22
			Invent	9	4.43
GROW UP	22	10.83	Mature	1	0.49
			Develop	0	0.00
GO ON	18	8.86	Continue	18	8.86
			Proceed	6	2.95
SET OUT	18	8.86	Explain	125	61.56
			Outline	11	5.42
			Arrange	4	1.97
OPEN UP	15	7.39	Reveal	7	3.45
			Uncover	7	3.45
			Unlock	3	1.48
TAKE ON	14	7.00	Engage	13	6.40
			Tackle	11	5.42
			Undertake	1	0.49
Total	**217**	**106.87**	***Total***	**1253**	**617.07**

*Per million words

The same pattern is observed here as for L2 students in that an inevitable preference is shown for PVs where no appropriate ALT exists (such as *end up* and *grow up*). *End up* has a relative frequency of 14.28, while the ALT *arrive* produced no results. Analysis of the PV *grow up* resulted in a relative

frequency of 10.83, while the ALT *mature* has a relative frequency of 0.49, and *develop* produced no results. (See §3.8.2.5 for further information on the ALT selection process.)

However, this is not the only pattern observed in cases of PV preference. For the PV *open up*, ALTs do exist (*reveal*, *uncover*, and *unlock*), but in this instance, the PV (with a relative frequency of 7.39) was evidently preferred to the ALTs (with relative frequencies of 3.45, 3.45 and 1.48, respectively). On the other hand, for the PV *take on*, with a relative value of 7.00, the preference is not so clear, as the ALT *engage* has a relative frequency of 6.40. In this case, the difference in use is comparatively minimal and does not suggest a decided preference for PVs over ALTS. Furthermore, for the PV *go on*, there appears to be no difference in use between it and the ALT *continue* (both having a relative frequency of 8.86). Given that there are instances where PV preference is only marginal, it is not surprising that the total values suggest a clear overall preference for ALTS over PVs (617.07 per million words for ALTS, compared to 106.87 for PVs), by a margin of 477.40 per cent.

4.2.1.2.4 Errors in NWU L1 PV use

Table 4.11 shows the PV errors that were found in the NWU L1 subcorpus. The most noteworthy error type, with an error type frequency of 2.95, was the arbitrary creation of a PV that does not exist, as in *break on*, *crack through*, *enter out*, *list down*, *stir out*, and *thrive off*. These were each used once only, except for *list down*, which appeared twice, hence the relative error token frequency of 3.45 for this error type.

Table 4.11: Errors in PV use in NWU L1 student writing

Type of PV error	***Error type frequency***	***Relative error type frequency****	***Error token frequency***	***Relative error token frequency****
Non-existent PV	6	2.95	7	3.45
Incorrect particle	1	0.49	2	0.98
Incorrect PV use	1	0.49	1	0.49
Total	**8**	**3.94**	**10**	**4.92**

*Per million words

The second error type, namely the addition of an incorrect particle, occurs only once, although the particular error appears twice. The PV *based of*, as found in the sentence *'…*content that is being*

*produced is **based of off**…'* is problematic for several reasons. It appears to be an incorrect version of *based off of*, found most often in colloquial American English, and, furthermore, is not recognised in any of the dictionaries used in this research. The correct PV to use in this context would be *based on*.

There is only one instance of a legitimate PV being used incorrectly, namely the PV *make out*. This is an informal PV with the meaning of dealing successfully with a situation or being intimate with someone (*Cambridge Phrasal Verbs Dictionary*, 2006). However, this is not the sense in which it is used in the subcorpus, as can be seen in the sentence: *'*These are just a few points **made out** by Matt Cavanagh…'*. Rather, in this case, it is likely that the student had intended to use either *set out* or, possibly, *made up*.

4.2.1.3 *Comparison of NWU L1 and L2 PV use*

In this subsection, the NWU L1 and L2 data extracted and discussed previously will be compared. The same parameters as used in the previous subsections will be applied here (PV type and token frequencies, PV error frequencies, distribution of PV data, and ten most frequently used PVs). The differences in patterns of PV use that emerge, if any, will be commented on.

4.2.1.3.1 Comparison of NWU L1 and L2 PV and error frequencies

Table 4.12, below, compares the total and relative PV frequency for NWU L1 and L2 students. There is a substantial difference in the token size of these subcorpora, with 2 030 568 tokens in the L1 subcorpus, and 14 898 759 tokens in the L2 subcorpus. PV use, therefore, appears to be much higher in the L2 subcorpus. However, when normalised per million words, the difference in PV use proves less than might have been supposed from the difference in subcorpora tokens, with a relative frequency of 268.89 for the L1 group, and a relative frequency of 287.07 for the L2 group. Error token frequency is also given in the table, and shows that, while PV use might be similar for the two groups, error frequency is more than three times as high in the L2 group than in the L1 group (*15.97 / 4.92 = 3.25*).

Table 4.12: Comparison of NWU L1 and L2 PV and error frequencies

Student group	Total token frequency	PV token frequency	Relative PV token frequency*	Error token frequency	Relative error token frequency*
L1 students	2 030 568	546	268.89	10	4.92
L2 students	14 898 759	4273	287.07	238	15.97

*Per million words

4.2.1.3.2 Comparison of PV distribution in NWU L1 and L2 subcorpora

Table 4.13 compares the total PV distribution in the NWU L1 and L2 subcorpora. When looking at the spread of relative PV token frequencies, the distribution between the two subcorpora present similar patterns, with the largest percentage of PV tokens (48.35% and 41.69%, respectively) being used five times or more, the next largest percentage (30.77% and 35.40%, respectively) falling within the middle category (between one and five occurrences per million words), and the smallest percentage of PVs appearing once or fewer times in the subcorpora (20.88% and 22.91% respectively).

Table 4.13: Comparison of PV distribution in NWU L1 and L2 student writing

Frequency group	L1				L2			
	PV tokens in this group	% of total PV tokens*	PV types represented	% of total PV types*	PV tokens in this group	% of total PV tokens*	PV types represented	% of total PV types*
Five or more occurrences per million words	130.01	48.35	6.89	9.85%	119.67	41.69	0.60	2.50%
Between one and five occurrences per million words	82.74	30.77	18.71	26.76%	101.62	35.40	3.49	14.50%
One or fewer occurrences per million words	56.14	20.88	44.31	63.39%	65.78	22.91	20.00	83.00%
Total	**268.89**	**100%**	**69.93**	**100%**	**287.07**	**100%**	**24.10**	**100%**

* Of normalised values

Table 4.13 also shows the distribution of PV types in the L1 and L2 subcorpora. The overall distribution pattern of the two groups again appears to be the same, with, in this case, the lowest percentage of use in the top category and the highest percentage in the bottom category. However, there are far fewer PV types in the top category of the L2 group (2.50%) than for the L1 group (9.85%). Likewise, far

more PV types occur in the bottom category of the L2 group (83%), than in the bottom category of the L1 group (63.39%). These results suggest that NWU L1 students have knowledge of and so are able to use a certain number of PVs appropriately on a regular basis, while L2 students are more inclined to use PVs incidentally.

4.2.1.3.3 Comparison of ten most frequently used PVs in NWU L1 and L2 subcorpora

Table 4.14 shows the comparison of the ten most frequently used PVs in the NWU L1 and L2 subcorpora. There are six PVs (indicated in light grey) that appear in both subcorpora, which suggests similarity in PV use by both groups. Interestingly, the PV *end up* appears to be the most frequently used PV in both cases, although its relative frequency differs substantially (with a relative frequency of 39.06 for the L2 group, and a relative frequency of 14.28 for the L1 group). The L2 frequencies are widely dispersed, with the difference between highest and lowest frequencies on the table being 34.23. Primarily, this difference appears to be as a result of the high frequency for *end up*. The dispersion of the L1 frequencies, with a difference of 7.28 between highest and lowest value, in contrast, is much lower than that of the L2 frequencies. The overall relative frequency of the L2 subcorpus is higher than that of the L1 subcorpus (124.5 to 106.96), which seems to support the higher L2 PV frequency reported in §4.2.1.3.1, although closer inspection shows that this is mainly driven by the high L2 frequency for *end up*.

Table 4.14: Comparison of NWU L1 and L2 PV use in student writing

L1 students			***L2 students***		
PV	***Absolute frequency***	***Relative frequency****	***PV***	***Absolute frequency***	***Relative frequency****
END UP	29	14.28	END UP	582	39.06
CARRY OUT	28	13.79	GROW UP	339	22.75
BRING ABOUT	25	12.31	CARRY OUT	160	10.74
COME UP	24	11.82	FIND OUT	154	10.34
MAKE UP	24	11.82	BRING ABOUT	134	8.99
GROW UP	22	10.83	GO ON	119	7.99
GO ON	18	8.86	SET OUT	100	6.71
SET OUT	18	8.86	STAND OUT	98	6.58
OPEN UP	15	7.39	GIVE OUT	97	6.51
TAKE ON	14	7.00	SEND OUT	72	4.83
Total	217	106.96	***Total***	1855	124.5

*Per million words

In this study, where appropriate, data are presented in a table and graph. The numbers in a table are easier to read and interpret, and provide overall totals which a graph does not. However, a graph provides a clear visual representation of differences. Furthermore, according to Brezina (2018:23), presenting data visually is useful in showing "the main trends in one's data". The box and whisker plot below (Figure 4.2) provides a visual representation of how the frequencies of the top ten PVs in the NWU L1 and L2 subcorpora compare. The L1 data show very narrow dispersion, with an interquartile range between 8.4925 and 12.68. The L1 ranges of the extreme value are small, further indicating narrow dispersion of data. The box of the L2 data also does not show wide dispersion, although the dispersion is greater than for the L1 data. The whiskers of a box-and-whisker plot (the lines that protrude from both sides of the box) indicate the low and high values or scores in the data set, with the low score on the bottom and the high score at the top. The ends of the two whiskers represent extreme values in a data set. If the whisker of one plot is longer than the whisker of another plot, it shows that the data in that data set are more scattered. In this case, the extreme values indicate that the L2 data are more scattered than the L1 data, and the data appear to be positively skewed. Furthermore, the L2 data contain an outlier, namely the PV *end up*, with a relative frequency of 39.06. In brief, the L2 data show wider dispersion than the L1 data, with the extension of the high score

whisker indicating that the extreme values are more scattered than the L1 values. Thus, the greater dispersion of L2 values compared to L1 values, as observed in the table, is confirmed here, showing high frequency of use at the top end of the table, rapidly diminishing towards the bottom end. This pattern of PV use suggests far less consistency in the use of PVs among L2 students.

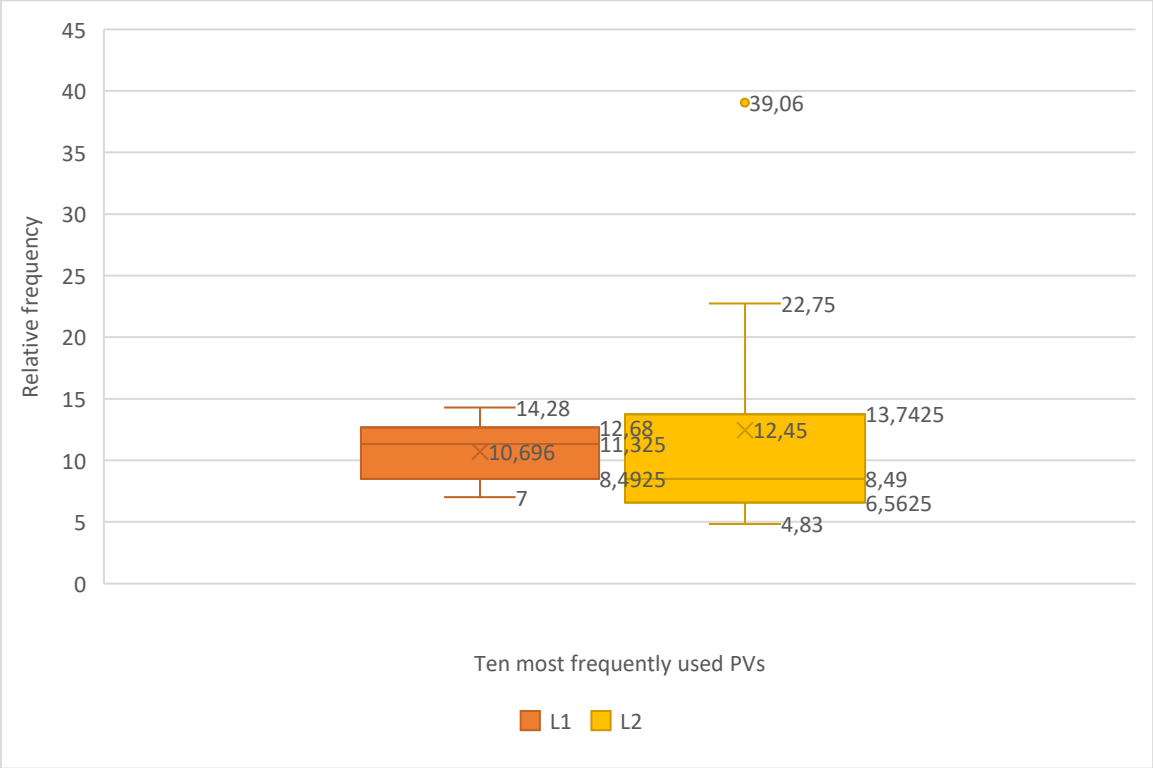

Figure 4.2: Distribution of NWU L1 and L2 top ten PVs

The NWU L1 and L2 PV use was examined in this section. This was the first of the three institutions to be investigated, with the UP and WITS data to follow

4.2.2 University of Pretoria (UP)

The second institution to be examined for its PV use is the University of Pretoria (UP). This is the smallest of the two SAMuLCAT subcorpora discussed in this research, with 6 747 522 tokens in total (2 033 648 L1 tokens and 4 713 874 L2 tokens). The L1 tokens make up 30% of the subcorpus, while the L2 tokens make up 70%. This information is presented in Table 4.15 below.

Table 4.15: UP L1 and L2 subcorpora		
Subcorpus	***Subcorpus size (tokens)***	***Percentage of total subcorpus size***
L1	2 033 648	30%
L2	4 713 874	70%
Total	**6 747 522**	**100%**

4.2.2.1 *UP L2 student PV use*

The UP L2 subcorpus produced 1 078 PVs in total, which equates to 228.69 PVs per million words (see Table 4.16 below). One-hundred and ninety-seven PV types, with a relative frequency of 41.79, are represented by these 1 078 PVs.

Table 4.16: Total PV frequency in UP L2 student writing		
Description	***Absolute frequency***	***Relative frequency****
PV types	197	41.79
PV tokens	1078	228.69

*Per million words

4.2.2.1.1 Distribution of PVs in the UP L2 subcorpus

Table 4.17 shows the distribution of PVs (per million words) in the UP L2 subcorpus. It would appear that most of the PV tokens (40.72%) were used between one and five times per million words, although only 20.81% of PV types (a relative frequency of 8.70) are represented in this group. A slightly smaller percentage (35.90%) of the total PVs occurred in the upper category (*five or more occurrences per million words*), although only a small percentage (4.06%) of PV types (a relative frequency of 1.70) are represented here. This suggests that students in this subcorpus have some knowledge of certain PVs, but that they are not in the habit of using particular PVs very often. On the other hand, the smallest percentage (23.38%) of PV tokens occurred in the lowest category (*one or fewer occurrences per million words*). This percentage is made up of 31.40 PV types per million words, or 75.13%, which is the highest proportion of PV types. Therefore, it seems that, although a small number of PVs are used more than once, most PVs in this subcorpus are used incidentally. Thus, it would appear that PV use is not a regular feature in UP L2 student writing.

Table 4.17: Distribution of total PVs per frequency group in UP L2 student writing

Frequency group	***Relative frequency* of PV tokens in this group***	***Relative frequency as percentage** of total PV tokens***	***Relative frequency* of PV types represented in this group***	***Relative frequency as percentage** of total PV types***
Five or more occurrences per million words	82.10	35.90%	1.70	4.06%
Between one and five occurrences per million words	93.13	40.72%	8.70	20.81%
One or fewer occurrences per million words	53.46	23.38%	31.40	75.13%
Total	**228.69**	**100%**	**41.79**	**100%**

*Per million words ** Of normalised values

4.2.2.1.2 Ten most frequently used PVs in the UP L2 subcorpus

Table 4.18 shows the ten most frequently used PVs in the UP L2 subcorpus. It is worth noting that the PV at the top of this list, namely *end up* (with 25.46 occurrences per million words), suggests a substantially higher frequency than is the case for any of the remaining PVs listed. The rapid diminishing of the PV frequencies is evidence that Zipf's law (see §4.2.1.1.2) is at work here. The dispersion of the data will be expanded on in §4.2.2.3.2.

Table 4.18: Ten most frequently used PVs in UP L2 student writing

PV	***Absolute frequency****	***Relative frequency****
END UP	120	25.46
GROW UP	77	16.33
POINT OUT	49	10.39
MAKE UP	32	6.79
CARRY OUT	31	6.58
BRING ABOUT	28	5.94
TAKE OVER	26	5.52
GO ON	24	5.09
BREAK DOWN	22	4.67
BRING UP	19	4.03

*Per million words

4.2.2.1.3 Use of ALTs in the UP L2 subcorpus

Table 4.19 shows ALTs for the ten most frequently used PVs in UP L2 student writing. The relative frequency of 602.90 for ALTs compared to the relative frequency of 90.80 for PVs indicates a clear overall preference for ALTs by UP L2 students. When breaking down these numbers, the following patterns can be observed: in cases where it is difficult to find suitable ALTs for PVs, as in the case of *end up*, *grow up* and *take over*, PV use is much higher than that of ALT use. In the other cases, three different scenarios are presented. Firstly, in the case of *go on* and *break down*, only the main ALT registered a higher frequency than the PV. In the second scenario, two of the ALTs (in the case of *carry out* and *bring up*) registered higher use than that of the PV. The PV *make up* could be said to fall either within scenario one or two, as its relative PV token frequency (6.79) is similar to that of the ALT *prepare* (6.36). In the third scenario, all three of the ALTs registered higher use than the PV, as occurred in *point out* and *bring about*. The analysis shows that, in each of these three scenarios, the ALTs accumulatively register higher frequencies than that of the PV.

Table 4.19: ALTs for the ten most frequently used PVs in UP L2 student writing

PV	***Absolute frequency***	***Relative frequency****	***ALT***	***Absolute frequency***	***Relative frequency****
END UP	120	25.46	Arrive	4	0.85
GROW UP	77	16.33	Develop	7	1.48
			Mature	2	0.42
POINT OUT	49	10.39	Mention	160	33.94
			Indicate	150	31.82
			Note	54	11.46
MAKE UP	32	6.79	Form	97	20.58
			Prepare	30	6.36
			Invent	7	1.48
CARRY OUT	31	6.58	Perform	64	13.58
			Accomplish	27	5.73
			Execute	23	4.88
BRING ABOUT	28	5.94	Cause	1122	238.02
			Produce	512	108.62
			Generate	117	24.82
TAKE OVER	26	5.52	Assume	3	0.64
			Appropriate	0	0.00
			Acquire	0	0.00
GO ON	24	5.09	Continue	110	23.34
			Persist	10	2.12
			Proceed	10	2.12
BREAK DOWN	22	4.67	Fail	48	10.18
			Collapse	11	2.33
			Decay	2	0.42
BRING UP	19	4.03	Raise	188	39.88
			Mention	76	16.12
			Educate	8	1.70
Total	**428**	**90.80**	***Total***	**2842**	**602.90**

*Per million words

4.2.2.1.4 Errors in UP L2 PV use

The errors in UP L2 PV use are shown in Table 4.20. The most prevalent type of error was the redundant particle, with a relative error type frequency of 4.45, and a relative error token frequency of 7.21. An example of this error in the subcorpus is found in the sentence: *'…*supporters of coercive*

state intervention argue that parents who fail to ***curb down*** *their children's obesity…'*, where the addition of the particle *down* to the verb *curb* is redundant.

Table 4.20: Errors in PV use in UP L2 student writing

Type of PV error	***Error type frequency***	***Relative error type frequency****	***Error token frequency***	***Relative error token frequency****
Redundant particle	21	4.45	34	7.21
Non-existent PV	15	3.18	15	3.18
Incorrect PV use	14	2.97	23	4.88
Incorrect particle	11	2.33	15	3.18
Unconfirmed PV	3	0.64	5	1.06
Total	**64**	**13.58**	**92**	**19.52**

*Per million words

The next error type is the use of a non-existent PV, with a relative PV error type frequency of 3.18. An example of this error can be found in the sentence, *'… *The economy of South Africa will continue to* ***grow down*** *as the due to unskilled and…*'. The PV *grow down* does not exist and has been created by the student, presumably to indicate an economic downturn. Three further examples of the 15 non-existent PVs used in this subcorpus are *air out* (*'*…assembled seminarians and church leaders nationwide to canvass ,* ***air out*** *and debate at length…*'), *bring through* (*'*… By simply generalizing the evidence* ***brought through*** *and the fact discussed above…')*, and *drip down* (*'*…their eyes swollen shut from all the chemicals that have been* ***dripped down*** *their eye sockets for hours on end….*').

The incorrect use of a PV has a relative error type frequency of 2.97. However, several of the error types occurred more than once, which resulted in a relative error token frequency of 4.88: *bring out* (4), *give out* (5), *put out* (2), and *put up* (2). An example of the incorrect use of a PV is seen in the use of *put up* in the sentence *'*…Seminars were supposed to be* ***put up*** *especially those that targeted…*' The PV *put up* is a legitimate PV but has been used incorrectly in this context.

The PV error type of attaching a wrong particle to a verb had a relative error type frequency of 2.33, and a relative error token frequency of 3.18. This discrepancy is again as a result of repeated use of particular errors, such as *make out*, *take on*, *bring down*. In these instances, while the PVs appear to

be legitimate, reference to the sentences in which they appear will illustrate that they are, in fact, incorrect. For example, the use of the PV, *bring about,* in the sentence: *'…*and the critics* ***bring about*** *valid points and reasoning*…', indicates incorrect use because of the addition of the particle *about* instead of *up*.

The presence of the last error type, the use of an unconfirmed PV, was minimal. This type of error had a relative error type frequency of 0.64, and a relative error token frequency of 1.06, because two of the PVs occurred twice in the subcorpus (*lock down* and *end off*). We again find a few cases where the PV that was used could not be confirmed, and yet seems to be in general use. An example of this would be *lock down*, a PV that has recently gained favour. The noun form, *lockdown*, is certainly familiar to us all, and the verb form is also in use, as can be seen in this BBC headline: "China Covid: Area around world's biggest iPhone plant locked down" (Liang, 2022). While this PV was not present in any of the reference works consulted at the time of this research, the PV might very well have been incorporated since then. The relative PV error type frequency in UP L2 student writing is 13.58, and the relative error token frequency is 19.52.

4.2.2.2 *UP L1 student PV use*

A total of 478 PV tokens, with a relative frequency 235.05, was found in the UP L1 subcorpus, as shown in Table 4.21 below. Within the token total, 143 PV types (with a relative frequency of 70.32) are represented.

Table 4.21: Total PV frequency in UP L1 student writing

Description	***Absolute frequency***	***Relative frequency****
PV types	143	70.32
PV tokens	478	235.05

*Per million words

4.2.2.2.1 Distribution of PVs in the UP L1 subcorpus

Table 4.22 shows the spread of PV types per million words in the UP L1 subcorpus. PV types that occur most frequently (five or more occurrences per million words) only represent 7.69% of all the PVs used.

PV types that occur between one and five times per million words are found three times as often in the subcorpus, but still represent only 23.08% of PV types present in the corpus. Most of the PV types (69.23%), therefore, occur only once or fewer times per million words.

Table 4.22: Distribution of PVs per frequency group in UP L1 student writing

Frequency group	***Relative frequency* of PV tokens in this group***	***Relative frequency as percentage** of total PV tokens***	***Relative frequency* of PV types represented in this group***	***Relative frequency as percentage** of total PV types***
Five or more occurrences per million words	93.43	39.75%	5.41	7.69%
Between one and five occurrences per million words	85.07	36.19%	16.23	23.08%
One or fewer occurrences per million words	56.55	24.06%	48.68	69.23%
Total	235.05	**100%**	70.32	**100%**

*Per million words ** Of normalised values

On the other hand, if we consider the PV tokens in this subcorpus, a different picture emerges. From this viewpoint, even though only a few PV types (5.41 per million words) occur five times or more, they occur often (39.75%). Likewise, the PVs that occur between one and five times per million words make up 36.19% of all the PVs used, though representing only 23.08% PV types. Furthermore, less than a quarter of the PV tokens in the subcorpus occur once or less per million words, although most of the PV types are represented here. Thus, what this comparison between tokens and types suggests is that a small percentage of PV types (7.69%) are used repeatedly (at least five times per million words), making up the largest percentage of PV tokens (39.75%). On the other hand, the largest percentage of PV types (69.23%) occur only once, thus making up the smallest percentage of PV tokens (24.06%). This is similar to the pattern noticed in previous subcorpora, with most PVs being used infrequently.

4.2.2.2.2 Ten most frequently used PVs in the UP L1 subcorpus

The most frequently used PVs are shown in Table 4.23. In this instance, there was a further PV that could have been added to the table, as *take over* had the same frequency of 11 occurrences (5.41

occurrences per million words) as the previous two PVs on the table (*build up* and *go on*). However, to keep the different subcorpora comparable, this PV was not included in the table, therefore, keeping the table to the top ten most frequently used PVs. The most frequently used PV, *end up*, with a relative PV token frequency of 15.74, was used substantially more than the next PV on the table, namely *bring about*, with a relative PV token frequency of 9.83. Zipf's law, though it might be said to apply to the first two frequencies, does not appear to be at work consistently in this subcorpus.

Table 4.23: Ten most frequently used PVs in UP L1 student writing

PV	***Absolute frequency (includes all versions of the base form of the PV)***	***Relative frequency****
END UP	32	15.74
BRING ABOUT	20	9.83
MAKE UP	20	9.83
POINT OUT	19	9.34
GROW UP	18	8.85
CARRY OUT	17	8.36
SLOW DOWN	16	7.87
FIND OUT	15	7.38
BUILD UP	11	5.41
GO ON	11	5.41

*Per million words

4.2.2.2.3 Use of ALTs in UP L1 subcorpus

Table 4.24 shows the ALTs for the ten most frequently used PVs in the UP L1 subcorpus. The highlighted (light grey) blocks indicate whether PVs or ALTs had the higher relative frequency. A comparison of frequencies strongly suggests that ALTs (with a relative frequency of 935.76) were preferred over PVs (with a relative frequency of 88.02).

Table 4.24: ALTs for the ten most frequently used PVs in UP L1 student writing

PV	*Absolute frequency*	*Relative frequency**	*ALT*	*Absolute frequency*	*Relative frequency**
END UP	32	15.74	Arrive	0	0
BRING ABOUT	20	9.83	Cause	437	214.88
			Produce	218	107.20
			Generate	64	31.47
MAKE UP	20	9.83	Form	51	25.08
			Prepare	9	4.43
			Invent	0	0.00
POINT OUT	19	9.34	Mention	86	42.29
			Indicate	79	38.85
			Note	52	25.57
GROW UP	18	8.85	Develop	8	3.93
			Mature	1	0.49
CARRY OUT	17	8.36	Perform	41	20.16
			Undertake	15	7.38
			Accomplish	8	3.93
SLOW DOWN	16	7.87	Impede	4	1.97
			Relax	3	1.48
			Decelerate	2	0.98
FIND OUT	15	7.38	Learn	52	25.57
			Detect	4	1.97
			Discover	2	0.98
BUILD UP	11	5.41	Increase	687	337.82
			Expand	16	7.87
			Multiply	7	3.44
GO ON	11	5.41	Continue	45	22.13
			Proceed	7	3.44
			Persist	5	2.46
Total	**179**	**88.02**	***Total***	**1903**	**935.76**

*Per million words

In the instances where the PV was preferred over the ALT (*end up, grow up*, and *slow down*), the low frequencies of the ALT suggest that the PV is used when no suitable alternative is found. In cases where the ALTs showed higher frequencies than the PVs (*bring about*, *make up*, *point out*, *carry out*, *find out*, *build up*, and *go on*), the differences were notable, showing clear preference for ALTs.

4.2.2.2.4 Errors in UP L1 PV use

Table 4.25 shows the errors in PV use in UP L1 student writing. The error type that occurs most often in this subcorpus is the incorrect use of a legitimate PV (with a relative PV error type frequency of 3.44). An example of this kind of error is found in this sentence: *'*These abilities will aid them in confronting the situations and not be* ***knocked down*** *by the outcomes…"'*, where, possibly, the intention was to indicate that "they" would be *bowled over,* or *overcome*, by the outcome. The relative error token frequency (3.44) is the same as the error type frequency, which means that no specific error example was repeated in the subcorpus.

Table 4.25: Errors in PV use in UP L1 student writing

Type of PV error	***Error type frequency***	***Relative error type frequency****	***Error token frequency***	***Relative error token frequency****
Incorrect PV use	7	3.44	7	3.44
Redundant particle	6	2.95	7	3.44
Incorrect particle	2	0.98	4	1.97
Unconfirmed PV	2	0.98	2	0.98
Total	**17**	**8.36**	**20**	**9.83**

*Per million words

The next most prevalent type of error is the unnecessary addition of a particle (with a relative error type frequency of 2.95), as in the sentence *'*…but the reality is dark days are upon us , if we do not* ***act up****…'. Clearly, the use of *act up* is incorrect here, as the behaviour referred to is neither that of a person behaving badly, nor that of a machine not working properly (*Cambridge Phrasal Verbs Dictionary*, 2006), but rather to *act* appropriately. Therefore, it can be deduced that the verb *act* should have been used here, and not the PV *act up*. In two instances, the particle *up* was incorrectly added to the verb *stand*, as in the sentence: *'*…us , as the residents of South Africa , should* ***stand up*** *behind our fellow citizens …'*, which resulted in the relative error token frequency being 3.44. (While *stand up* is a valid PV, it clearly was not used correctly in this context).

The last two types of errors both have a relative error type frequency of 0.98, though the first of the two types of error has a relative PV error token frequency of 1.97. The first type of error refers to an

incorrect particle being connected to a PV. *Cut off* is a legitimate PV when used correctly, as in the sentence '*… rhinos are darted and put under anaesthesia , and their horns are **cut off** with chainsaws …*', but in the sentence *'***Cutting off** *certain foods might help combat cancer…'* the wrong particle has clearly been combined with the verb *cut*. In this case, the correct particle would be *out* (as in *cut out*). This particular error (*cut off*) occurred three times, which explains the discrepancy between the relative error type frequency and the relative error token frequency.

The final error type is where a PV cannot be validated, but appears to be legitimate. An example of this can be seen in the sentence *'*Prayer requests may be **sent through** for people to be prayed for…*'. Interestingly, although the PV could not be found in any dictionary, it is referred to online on several websites of an informal nature (Ludwig.guru, 2024; StackExchange, 2024; WordReference.com, 2024). This suggests that the PV *send through* is being used by the general public, and is, therefore, likely to be recognised in the near future as a PV. The same is also true for *carry across*, the other PV error of this type (*Word Hippo*, 2008; *Glosbe Multiligual Online Dictionary*, 2024; Reverso, 2024).

An interesting observation is the habit displayed by the students in this subcorpus of hyphenating the PV, as can be seen in the sentence: *'*The essay will then **sum-up** what criteria need to be met…*'. Other examples in this subcorpus include *phase-out* (as in *'*…many world leaders accepted to help **phase-out** substances …*'), *start-up* (as in *'*…they can't afford the funding to **start-up** an alternative source of energy …*'), *wake-up* (as in *'*…Humans need to **wake-up** and make a choice…*'). This suggests that the students understand the unity inherent in a PV, although apparently without the realisation that the process of inserting a hyphen between verb and particle can potentially change the part of speech. These instances were not indicated as PV errors, as they were considered to fall within the sphere of punctuation. Nevertheless, it was deemed worthwhile to comment on this anomaly, especially as it was observed in L1 student writing.

4.2.2.3 *Comparison of UP L1 and L2 PV use*

While the L1 and L2 PV use in the UP subcorpus was observed and commented on in the previous subsections, it is also necessary to understand how these two groups relate to each other. As before, PV type and token frequencies, PV error frequencies, PV type and token distribution, and most frequently used PVs will be the parameters used to observe differences between the two groups.

4.2.2.3.1 Comparison of UP L1 and L2 PV and error frequencies

Table 4.26 confirms that overall PV use was similar for UP L1 and L2 students (with relative PV token frequencies of 235.05 and 228.69, respectively). However, the L2 subcorpus had close to twice as many errors as the L1 subcorpus (a relative PV error frequency of 19.52, compared to a relative PV error frequency of 9.83).

Table 4.26: Comparison of UP L1 and L2 PV and error frequencies

Student group	***Total token frequency***	***PV token frequency***	***Relative PV frequency****	***Error token frequency***	***Relative error token frequency****
L1 students	2 033 648	478	235.05	20	9.83
L2 students	4 713 874	1 078	228.69	92	19.52

*Per million words

4.2.2.3.2 Comparison of PV distribution in UP L1 and L2 subcorpora

A comparison of the spread of PV types per million words in UP student writing is given in Table 4.27. A similar pattern of use is evident in both subcorpora, in that the largest percentage of PV type use occurred at the lowest level (one or fewer occurrences per million words), and the lowest percentage of PV type use at the highest level (five or more occurrences per million words). However, the distribution is different for the two subcorpora. L1 students used more PV types that occurred five times or more per million words than did the L2 students (7.69% and 4.06% respectively, thus a percentage point difference of 3.63). L1 use of PV types is also slightly higher in the middle category (between one and five occurrences per million words), with 23.08%, compared to a 20.81% L2 use of PV types (a difference of 2.27 percentage points). On the other hand, the L2 students used more PV

tokens that occurred only once or less per million words (69.23% and 75.13% respectively, thus a difference of 5.9 percentage points). The L1 students, therefore, show more familiarity with PVs than the L2 students, in that they are more inclined to use certain PVs consistently, while L2 students are more inclined to use PVs once only, suggesting incidental use rather than knowledge of a select range of PVs. In total, the L1 students used more PV types than the L2 students (with a relative PV type frequency of 70.32 for L1 compared to a relative PV type frequency of 41.79 for L2). It can, therefore, be concluded that the L1 students used a greater variety of PVs.

Table 4.27: Comparison of PV distribution in UP L1 and L2 student writing

Frequency group	***L1***				***L2***			
	PV tokens in this group	***% of total PV tokens****	***PV types represented***	***% of total PV types****	***PV tokens in this group***	***% of total PV tokens****	***PV types represented***	***% of total PV types****
Five or more occurrences per million words	93.43	39.75%	5.41	7.69%	82.10	35.90%	1.70	4.06%
Between one and five occurrences per million words	85.07	36.19%	16.23	23.08%	93.13	40.72%	8.70	20.81%
One or fewer occurrences per million words	56.55	24.06%	48.68	69.23%	53.46	23.38%	31.40	75.13%
Total	235.05	**100%**	70.32	**100%**	**228.69**	**100%**	**41.79**	**100%**

* Of normalised values

Table 4.27 also gives the comparison of PV token distribution per million words in UP L1 and L2 student writing. When doing a comparison of the PV tokens generated by both groups, we see that most of the PV tokens in the L1 group occurred five or more times per million words, whereas most of the PV tokens in the L2 group occurred between one and five times per million words. In both groups, the smallest percentage of PV tokens (24.06% for the L1 group, and 23.38% for the L2 group) occurred in the lowest category (one or fewer occurrences per million words). The overall picture for both subcorpora shows that a small number of PV types are used at high frequencies, whereas the bulk of the PV types seems to be used by chance, suggesting purely incidental use of PVs. The number of PV types used most frequently by the L2 students is particularly small, representing only 4.06% of the total number used.

4.2.2.3.3 Comparison of ten most frequently used PVs in UP L1 and L2 subcorpus

Table 4.28 shows the comparison between the L1 and L2 student PV use in the UP corpus. Seven of the ten most frequently used PVs are featured in both subcorpora, although at different frequencies. The total use of high-frequency PVs is similar in the two subcorpora (with a relative frequency of 88.02 for the L1 subcorpus, and a relative frequency of 90.80 for the L2 subcorpus).

Table 4.28: Comparison of UP L1 and L2 PV use in student writing

L1 students			***L2 students***		
PV	***Absolute frequency***	***Relative frequency****	***PV***	***Absolute frequency***	***Relative frequency****
END UP	32	15.74	END UP	120	25.46
BRING ABOUT	20	9.83	GROW UP	77	16.33
MAKE UP	20	9.83	POINT OUT	49	10.39
POINT OUT	19	9.34	MAKE UP	32	6.79
GROW UP	18	8.85	CARRY OUT	31	6.58
CARRY OUT	17	8.36	BRING ABOUT	28	5.94
SLOW DOWN	16	7.87	TAKE OVER	26	5.52
FIND OUT	15	7.38	GO ON	24	5.09
BUILD UP	11	5.41	BREAK DOWN	22	4.67
GO ON	11	5.41	BRING UP	19	4.03
Total	**179**	**88.02**	***Total***	**428**	**90.80**

*Per million words

The box-and-whisker graph in Figure 4.3 illustrates the dispersion of the L1 and L2 subcorpora data. The difference in sizes of the boxes shows that the L1 data are less dispersed than that of the L2 data. Fifty per cent of the L1 data (interquartile range) range from 6.8875 to 9.83, with 9.83 also being the maximum value. The median value is close to the mean value, which suggests normal distribution. On the other hand, the L2 box is more extended than the L1 box, showing that the L2 data set is more widely dispersed than the L1 data set. With a median value of 6.26 and a mean value of 9.08, the L2 data show right-leaning abnormal dispersion. The maximum value indicates that the high scores are more widely distributed than that of the L1 values. Both subcorpora have an outlier, and in both cases, it is the PV *end up* (with a relative frequency of 15.74 in the L1 subcorpus, and a relative frequency of

25.46 in the L2 subcorpus). Therefore, although the total relative frequencies for the ten most used PVs are similar for the two subcorpora, their data are differently dispersed.

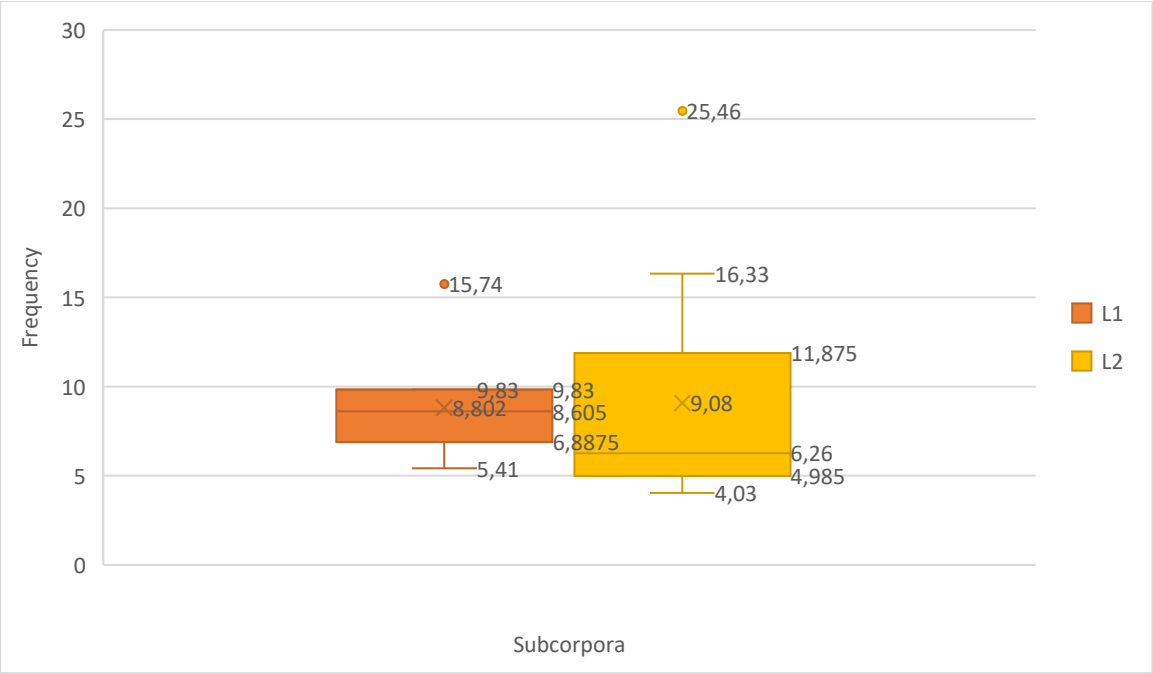

Figure 4.3: Distribution of UP L1 and L2 top ten PVs

In this subsection, PV use in the UP subcorpus, the second subcorpus in the SAMuLCAT corpus, was discussed. L1 and L2 PV use was again assessed separately, before doing a comparison of the two groups. The final institution, WITS, will now be discussed.

4.2.3 University of the Witwatersrand (WITS)

The third institution to be used in the study is the University of the Witwatersrand. This corpus is not one of the SAMuLCAT subcorpora, but rather a corpus on its own (see §3.8.1.2). It consists of 5 087 818 L1 tokens and 3 336 391 L2 tokens, giving a total of 8 424 209 tokens, as shown in Table 4.29 below. The L1 tokens make up 60% of the corpus, and the L2 tokens make up 40%. This is the only institution of the three used in this study where the L1 tokens are higher than the L2 tokens.

Table 4.29: WITS L1 and L2 subcorpora		
Subcorpus	***Subcorpus size (tokens)***	***Percentage of total subcorpus size***
L1	5 087 818	60%
L2	3 336 391	40%
Total	**8 424 209**	**100%**

4.2.3.1 *WITS L2 student PV use*

As indicated in Table 4.30 below, 3 120 PV tokens were found in the WITS L2 subcorpus, which equates to a relative frequency of 935.14 PVs. The tokens are made up of 315 PV types, with a relative frequency of 94.41.

Table 4.30: Total PV frequency in WITS L2 student writing		
Description	***Absolute frequency***	***Relative frequency****
PV types	315	94.41
PV tokens	3 120	935.14

*Per million words

4.2.3.1.1 Distribution of PVs in the WITS L2 subcorpus

Table 4.31 shows the distribution of PVs in the WITS subcorpus. While 66.47% of total PVs appear five times or more per million words, these PVs represent only 10.79% of the PV types present in the subcorpus, indicating that these particular PVs have a notably high use. On the other hand, while the largest percentage of PV types (58.73%) appears once or less per million words, they represent only 9.30% of PV token use. This suggests that 41.27% of PV types are used more than once, with 10.79% of these being used with the greatest frequency. However, most PV types in this subcorpus (58.73%) are used once only, suggesting that they are used by chance, without substantial knowledge of the PVs.

Table 4.31: Distribution of PVs per frequency group in WITS L2 student writing

Frequency group	***Relative frequency* of PV tokens in this group***	***Relative frequency as percentage** of total PV tokens***	***Relative frequency* of PV types represented in this group***	***Relative frequency as percentage** of total PV types***
Five or more occurrences per million words	621.63	66.47%	10.19	10.79%
Between one and five occurrences per million words	226.59	24.23%	28.77	30.48%
One or fewer occurrences per million words	86.92	9.30%	55.45	58.73%
Total	935.14	**100%**	94.41	**100%**

*Per million words * Of normalised values

4.2.3.1.2 Ten most frequently used PVs in the WITS L2 subcorpus

The ten most frequently used PVs are shown in Table 4.32. Elements of Zipf's law are evident, but not on a consistent basis. The dispersion of the data presented here will be discussed in more detail when the WITS L1 and L2 data are compared.

Table 4.32: Ten most frequently used PVs in WITS L2 student writing

PV	***Absolute frequency***	***Relative frequency****
GROW UP	411	123.19
CARRY OUT	217	65.04
MAKE UP	201	60.24
END UP	160	47.96
FIND OUT	124	37.17
PICK UP	75	22.48
POINT OUT	57	17.08
COME UP	52	15.59
OPEN UP	48	14.39
TURN OUT	48	14.39

*Per million words

4.2.3.1.3 Use of ALTs in the WITS L2 subcorpus

The ALTs for the ten PVs used most frequently by WITS L2 students are given in Table 4.33. The method for choosing ALTs was the same as for the previous subcorpora. In some cases, several ALTs had to be checked before viable alternatives could be found. For example, for the PV *turn out*, the verbs

extinguish, *happen* and *expel* produced no results. Subsequently, *produce, emerge* and *transpire* were checked for viability and proved more useful.

Table 4.33: ALTs for the ten most frequently used PVs in WITS L2 student writing

PV	***Absolute frequency***	***Relative frequency****	***ALT***	***Absolute frequency***	***Relative frequency****
GROW UP	411	123.19	Mature	33	9.89
			Develop	53	15.89
CARRY OUT	217	65.04	Perform	419	125.58
			Implement	152	45.56
			Accomplish	72	21.58
MAKE UP	201	60.24	Create	585	175.34
			Form	505	151.36
			Constitute	79	23.68
END UP	160	47.96	Arrive	24	7.19
FIND OUT	124	37.17	Learn	592	177.44
			Discover	160	47.96
			Detect	49	14.69
PICK UP	75	22.48	Collect	35	10.49
			Lift	7	2.10
			Fetch	2	0.60
POINT OUT	57	17.08	Mention	455	136.37
			Indicate	276	82.72
			Highlight	210	62.94
COME UP	52	15.59	Occur	1175	352.18
			Arise	197	59.05
			Emerge	116	34.77
OPEN UP	48	14.39	Reveal	80	23.98
			Disclose	55	16.48
			Uncover	16	4.80
TURN OUT	48	14.39	Produce	213	63.84
			Emerge	48	14.39
			Transpire	1	0.30
Totals	**1 393**	**417.53**	***Totals***	**5 609**	**1681.17**

*Per million words

Three of the PVs (namely *grow up*, *end up* and *pick up*) showed a higher frequency than the ALTs (indicated in light grey). These three PVs fall within the category of PVs for which no convenient

synonyms exist, as is evident from the limited choice of ALTs for *grow up* (two only: *mature* and *develop*), and *end up* (one only: *arrive*).

In the case of the seven remaining PVs (*carry out*, *make up*, *find out*, *point out*, *come up*, *open up*, and *turn out*), ALTs were not only preferred, but substantially so. Consequently, for the ten PVs considered here, the total frequency is higher for the ALTs (with a relative frequency of 1681.17) than for the PVs (with a relative frequency 417.53) by 302.65% ((1681.17 – 417.53) / 417.53 * 100).

4.2.3.1.4 Errors in WITS L2 PV use

WITS L2 PV errors are shown in Table 4.34. All five error types were found in this subcorpus. Two of the error types were more prevalent than the other three by a substantial margin. The first of these (with a relative error type frequency of 10.49, and a relative error token frequency of 13.79) was the adding of a redundant particle to the verb proper. An example of this error is found in the following two sentences. Firstly, in the sentence, *'…*they have strengths and weaknesses as the world is open to all that **happens around**. The theory…*', the particle *around* is an unnecessary addition to the verb *happens*. In the second sentence, *'…*To do this duties , they have to **offer out** tenders to people…*', the particle *out* has been added unnecessarily to the verb *offer*.

Table 4.34: Errors in PV use in WITS L2 student writing

Type of PV error	***Error type frequency***	***Relative error type frequency****	***Error token frequency***	***Relative error token frequency****
Redundant particle	35	10.49	46	13.79
Incorrect PV use	31	9.29	49	14.69
Incorrect particle	12	3.60	15	4.50
Non-existent PV	7	2.10	9	2.70
Unconfirmed PV	7	2.10	7	2.10
Total	**92**	**27.58**	**126**	**37.78**

*Per million words

The second most prevalent type of error (with a relative error type frequency of 9.29, and a relative error token frequency of 14.69) is the incorrect use of a legitimate PV. An example of this type of error

is found in the sentence, *'…*are permeable in the sense that the work responsibilities can easily* ***brim over*** *into the family sphere*…'. While *brim over* is a legitimate PV, its use in this sentence is not. A PV such as *spill over* would have been a better choice here.

Adding an incorrect particle to the verb proper was the next most frequent type of error, with a relative error type frequency of 3.60, and a relative error token frequency of 4.50. An example of this type of error is *'…*therefore in spite of the ethnicity diversity , one can not* ***rule off*** *the fact not every South African is culturally influenced*…', where context suggests that the use of the PV *rule out* would have been the better choice.

The fourth type of error found was that of the use of a non-existent PV (with a relative error type frequency of 2.10, and an error token frequency of 2.70). An example is *'*Violent patterns in South Africa date back to pre-colonial times ,* ***stretching over*** *to the apartheid regime*…'. Possibly, the student had meant to say, '***crossing over*** *to the apartheid regime'*.

For the fifth and last error, the use of an unconfirmed PV (with a relative error type frequency of 2.10, and, likewise, a relative error token frequency of 2.10), an example is '… *Bud valued friendship with his colleagues and they often* ***joke around*** *and are jolly together*…'. While this PV could not be confirmed officially, it is referenced informally online (*The Free Dictionary*, 2024; *Wiktionary*, 2024; *Wordnik*, 2024; *YourDictionary*, 2024), which suggests that its use is not unfamiliar to the public.

Overall, in the case of the WITS L2 students, the error types totalled 92 (with a relative error type frequency of 27.58), and the total error token frequency came to 126 (with a relative error token frequency of 37.78).

4.2.3.2 *WITS L1 student PV use*

The WITS L1 subcorpus delivered a total of 3 933 legitimate PVs (which equates to a relative PV token frequency of 773.02) (see Table 4.35 below). From these tokens, 361 PV types were identified (with a relative PV type frequency of 70.95).

Table 4.35: Total PV frequency in WITS L1 student writing		
Description	***Absolute frequency***	***Relative frequency****
PV types	361	70.95
PV tokens	3 933	773.02

*Per million words

1..1.1.1 Distribution of PVs in the WITS L1 subcorpus

Table 4.36 gives the distribution of PVs in the WITS L1 subcorpus. It appears that most of the PV tokens (61.73%) occur five time or more per million words. On the other hand, only 7.76% of PV types are represented in this group, which suggests that there is a high level of familiarity with these PV types, since each PV is used 87 times on average. Far more PV types (23.55%) occur between one and five times per million words, and these PVs make up about a quarter of total PV tokens used (25.02%). Furthermore, although only 13.25% of the PV tokens are used once only, this represents 68.69% of the PV types in this subcorpus. Therefore, while most PVs in this subcorpus are used once only, suggesting incidental use rather than knowledge of the PVs used, the PVs that have been used at least five times show high frequency of use.

Table 4.36: Distribution of PVs per frequency group in WITS L1 student writing				
Frequency group	***Relative frequency* of PV tokens in this group***	***Relative frequency as percentage** of total PV tokens***	***Relative frequency* of PV types represented in this group***	***Relative frequency as percentage** of total PV types***
Five or more occurrences per million words	477.22	61.73%	5.50	7.76%
Between one and five occurrences per million words	193.40	25.02%	16.71	23.55%
One or fewer occurrences per million words	102.40	13.25%	48.74	68.69%
Total	**773.02**	**100%**	**70.95**	**100%**

*Per million words ** Of normalised values

4.2.3.2.1 Ten most frequently used PVs in the WITS L1 subcorpus

Table 4.37 shows the ten most frequently used PVs in this corpus. Zipf's law is not in evidence here, as, apart from the 30.86 drop from 57.98 (*grow up*) to 27.12 (*act out*), there is no evidence of the word

frequency dropping rapidly. The distribution of the data will be discussed in more detail when the WITS L1 and L2 data are compared.

Table 4.37: Ten most frequently used PVs in WITS L1 student writing

PV	***Absolute frequency***	***Relative frequency****
CARRY OUT	325	63.88
MAKE UP	303	59.55
GROW UP	295	57.98
ACT OUT	138	27.12
BRING UP	133	26.14
PICK UP	113	22.21
FIND OUT	108	21.23
OPEN UP	80	15.72
SET OUT	80	15.72
POINT OUT	77	15.13

*Per million words

4.2.3.2.2 Use of ALTs in the WITS L1 subcorpus

Table 4.38 gives a comparison of the ten most frequently used PVs and their ALTs in the WITS L1 subcorpus. It appears that, for seven of the PVs (*carry out*, *make up*, *bring up*, *find out*, *open up*, *set out* and *point out*), ALTs were preferred. On the other hand, the use of the PV was preferred in three instances (*grow up*, *act out* and *pick up*). The PVs *grow up* and *act out* fall within the category of PVs for which there is no suitable alternative.

Table 4.38: ALTs for the ten most frequently used PVs in WITS L1 student writing

PV	***Absolute frequency***	***Relative frequency****	***ALT***	***Absolute frequency***	***Relative frequency****
CARRY OUT	325	63.88	Perform	858	168.64
			Conduct	822	161.56
			Implement	263	51.69
MAKE UP	303	59.55	Create	1248	245.29
			Form	896	176.11
			Constitute	98	19.26
GROW UP	295	57.98	Develop	108	21.23
			Mature	50	9.83
ACT OUT	138	27.12	Demonstrate	109	21.42
			Illustrate	82	16.12
			Represent	70	13.76
BRING UP	133	26.14	Mention	553	108.69
			Raise	273	53.66
			Educate	15	2.95
PICK UP	113	22.21	Collect	82	16.12
			Fetch	6	1.18
			Lift	3	0.59
FIND OUT	108	21.23	Learn	897	176.30
			Discover	345	67.81
			Detect	31	6.09
OPEN UP	80	15.72	Reveal	206	40.49
			Discover	188	36.95
			Uncover	29	5.70
SET OUT	80	15.72	Explain	925	181.81
			Outline	123	24.18
			Arrange	36	7.08
POINT OUT	77	15.13	Indicate	614	120.68
			Note	569	111.84
			Highlight	357	70.17
Total	**1 652**	**324.68**	***Total***	**9 856**	**1 937.20**

*Per million words

This apparent preference for ALTs is confirmed when considering the total frequencies. The total relative PV frequency is 324.68, whereas the total relative ALT frequency is 1 937.20, which is 496.65% higher than the PV frequency. There is thus a marked preference for ALTs by the WITS L1 students represented in this subcorpus.

4.2.3.2.3 Errors in WITS L1 PV use

Table 4.39 gives the errors found in WITS L1 student writing. The error that occurred most was the unnecessary addition of a particle, with a relative error type frequency of 7.86, and a relative error token frequency of 11.79. An example of the unnecessary addition of a particle to the verb proper is *'*…businesses, homes and apartments were* ***bought over*** *by many foreign nationals and South Africans*…' where the particle *over* has unnecessarily been added to the verb *bought*. In the following example of the same error, the redundancy of the particle only emerges after careful examination of the meaning of the sentence: *'*…and in his attention to deal [possibly* ***detail****?] when* ***dressing up*** *after the tennis game*…'. *Dressing up* suggests preparing for an elaborate affair, whereas the sentence simply seems to refer to changing the subject's clothes after having participated in a sporting event.

Table 4.39: Errors in PV use in WITS L1 student writing

Type of PV error	***Error type frequency***	***Relative error type frequency****	***Error token frequency***	***Relative error token frequency****
Redundant particle	40	7.86	60	11.79
Incorrect PV use	17	3.34	25	4.91
Non-existent PV	16	3.14	23	4.52
Unconfirmed PV	11	2.16	20	3.93
Incorrect particle	11	2.16	19	3.73
Total	**95**	**18.67**	**147**	**28.89**

*Per million words

The use of a redundant particle in the formation of a PV occurred with more than double the frequency of the next two errors listed in the table, that of the incorrect use of a legitimate PV and the use of a non-existent PV, both of which had a similar relative PV error type frequency. The first of these error types had a relative PV error type frequency of 3.34, and a relative PV error token frequency of 4.91,

while the use of a non-existent PV had a relative PV error type frequency of 3.14, and a relative PV error token frequency of 4.52. An example of the first error is *'… *some adolescence do in fact* ***come out*** *to be resilient*…'. The PV *come out* is legitimate; however, its use in this sentence is incorrect. The student probably intended to use the PV *turn out*. Secondly, the sentence *'*…she mentally escapes to a fantasy whereby she is attending a movie premiere ,* ***primed up*** *in a red dress*…' contains the illegitimate PV *prime up*.

The last two error types also have a similar incidence per million words, with a relative PV error type of 2.19, although there is a slight variation as far as relative PV error token is concerned. The first of these errors (with a relative PV error token frequency of 3.93), is the use of unconfirmed PVs which, while their legitimacy cannot be confirmed, are referred to informally online. An example of this error is '… *When she was on the floor she tried to* ***crawl away*** *looking terribly fearful*…". The PV *crawl away*, while not included in any of the sources identified as containing standard or widely recognised PVs, is referenced by several informal online sites, such as *Glosbe Multiligual Online Dictionary* (2024), *Linguee* (2024) and *Thesaurus.plus* (2024). The second error (with a relative PV error token frequency of 3.73) is an incorrect particle being attached to a verb, such as in the sentence: *'*…essentially* ***weigh out*** *the costs and benefits of smoking*… ', where the correct PV combination would have been *weigh up*.

4.2.3.3 *Comparison of WITS L1 and L2 PV use*

WITS L1 and L2 PV patterns of use have been observed and commented on in the previous subsections. In this subsection, the two groups will be compared in order to establish how they differ, if at all. By this means, a more rounded picture of the subcorpus as a whole should emerge. As before, the parameters that will be used for the comparison are PV type and token frequencies, PV error frequencies, PV distribution, and ten most used PVs.

4.2.3.3.1 Comparison of WITS L1 and L2 PV and error frequencies

Table 4.40 provides a comparison of the L1 and L2 PV type and token frequencies in the WITS corpus, as well as the PV error frequencies. This is the only institution where the L1 group is larger than the L2 group. It is, therefore, notable that the L2 group shows the higher use of PVs, with a relative frequency of 935.14, compared to the L1 relative frequency of 773.02. In line with the findings of the NWU and UP data, the relative PV error tokens are higher for the L2 group than for the L1 group. Thus, it can be surmised that, while the L2 students use more PVs, they are also more likely to make mistakes when using them.

Table 4.40: Comparison of WITS L1 and L2 PV and error frequencies

Student group	***Total token frequency***	***PV token frequency***	***Relative PV token frequency****	***Error token frequency***	***Relative error token frequency****
L1 students	5 087 818	3 933	773.02	147	28.89
L2 students	3 336 391	3 120	935.14	126	37.78

*Per million words

4.2.3.3.2 Comparison of distribution of PVs in WITS L1 and L2 subcorpus

Table 4.41 shows the comparison of total PV distribution per million words between WITS L1 and L2 students. The L2 subcorpus has a higher percentage of PV tokens (66.47%) in the category of five or more occurrences per million words, compared to the L1 subcorpus (61.73%), a difference of 4.74 percentage points. While there is a similarity in PV token frequency in the middle category (between one and five occurrences per million words) of the two subcorpora, a difference can again be noted in the lower category (one or fewer occurrences per million words). Here, the L1 token frequency (13.25%) is 3.95 percentage points higher than the L2 token frequency (9.30%), which suggests that the WITS L1 students are more likely than the L2 students to use PVs once only. However, the overall picture here is that both groups of students generally used PVs five times or more.

Table 4.41: Comparison of PV distribution in WITS L1 and L2 student writing

Frequency group	***L1***				***L2***			
	PV tokens in this group	***% of total PV tokens****	***PV types represented***	***% of total PV types****	***PV tokens in this group***	***% of total PV tokens****	***PV types represented***	***% of total PV types****
Five or more occurrences per million words	477.22	61.73	5.50	7.76	621.63	66.47	10.19	10.79
Between one and five occurrences per million words	193.40	25.02	16.71	23.55	226.59	24.23	28.77	30.48
One or fewer occurrences per million words	102.40	13.25	48.74	68.69	86.92	9.30	55.45	58.73
Total	773.02	100%	70.95	100%	935.14	100%	94.41	100%

* Of normalised values

Apart from addressing the spread of PV tokens in the two subcorpora, Table 4.41 also shows the spread of PV types in the two subcorpora. It seems that most PV types fall in the bottom category (one or fewer occurrences per million words) for both groups, although the L1 subcorpus has a higher percentage of PV types in this category (68.69%), compared to L2 students (58.73%). On the other hand, the L2 group has higher percentages of PV type use in both the other categories (30.48% between one and five occurrences, compared to the 23.55% of the L1 group); and 10.79% for five or more occurrences, compared to the 7.76% of the L1 group. This suggests that L2 students are more inclined to use specific PVs more than once.

In summary, the PV type and token information in the table suggests that students from both corpora generally use PVs only once. However, the L2 students seem to be more likely than the L1 students to use a PV more than once. Both subcorpora display frequent use of a small percentage of PVs, although the percentage is larger for the L2 subcorpus than for the L1 subcorpus. Data distribution will be discussed further in the next subsection (§4.2.3.3.3).

4.2.3.3.3 Comparison of ten most frequently used PVs in WITS L1 and L2 subcorpora

Table 4.42 compares the ten most frequently used PVs by WITS L1 and L2 student groups. The two groups show similarity in PV use, in that the same three PVs (*carry out*, *make up* and *grow up*) appear at the top of both lists (indicated in light grey), although at different frequencies. Four further PVs are shared (*pick up*, *open up*, *find out* and *point out*). Overall, PV use is higher in the L2 subcorpus (with a relative frequency of 417.53) than in the L1 subcorpus (with a relative frequency of 324.68). This seems to confirm the pattern of PV use observed in the previous part of the research (§4.2.3.3.2), which indicated that the L2 subcorpus shows a higher overall PV frequency than the L1 subcorpus (especially in the case of PVs that occur five or more times per million words).

Table 4.42: Comparison of WITS L1 and L2 PV use in student writing

L1 students			***L2 students***		
PV	***Absolute frequency***	***Relative frequency****	***PV***	***Absolute frequency***	***Relative frequency****
CARRY OUT	325	63.88	GROW UP	411	123.19
MAKE UP	303	59.55	CARRY OUT	217	65.04
GROW UP	295	57.98	MAKE UP	201	60.24
ACT OUT	138	27.12	END UP	160	47.96
BRING UP	133	26.14	FIND OUT	124	37.17
PICK UP	113	22.21	PICK UP	75	22.48
FIND OUT	108	21.23	POINT OUT	57	17.08
OPEN UP	80	15.72	COME UP	52	15.59
SET OUT	80	15.72	OPEN UP	48	14.39
POINT OUT	77	15.13	TURN OUT	48	14.39
Total	**1 652**	**324.68**	***Total***	**1 393**	**417.53**

*Per million words

The information in Table 4.42 is represented in a box and whisker plot in Figure 4.4. The interquartile ranges of both data sets are similarly distributed, and both are right skewed, which means that most of the values are at the lower end of the distribution, indicating that a greater proportion of the PVs occur at lower frequencies than the median. However, the extended whisker (high score) of the L2 data set indicates that the L2 values are more scattered at the higher end than for the L1 data set.

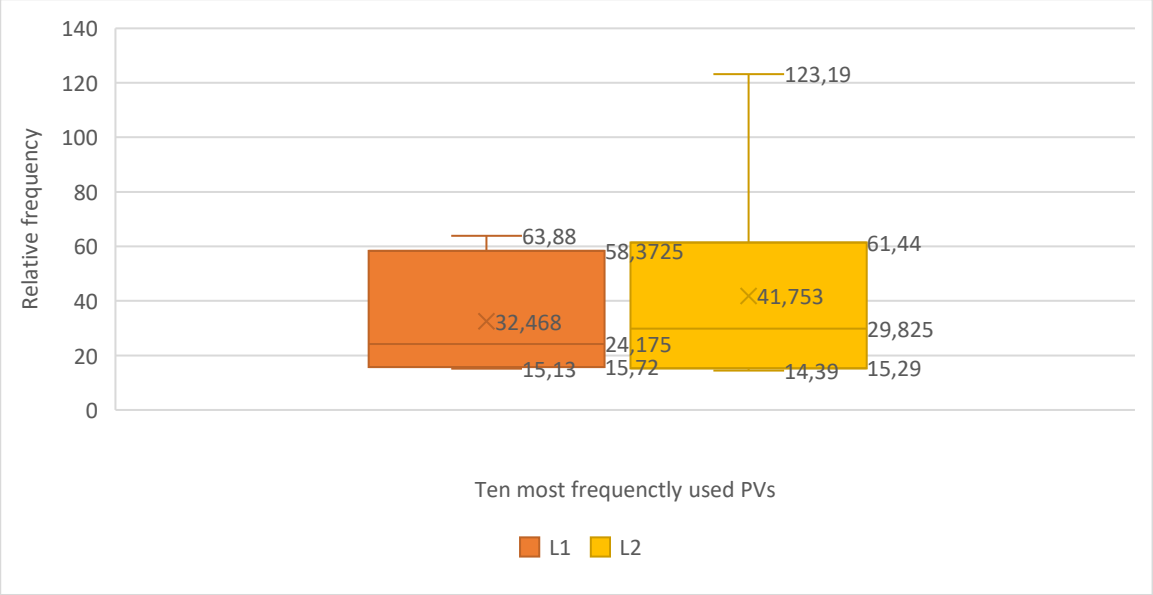

Figure 4.4: Distribution of WITS L1 and L2 top ten PVs

Therefore, as was noticed in the previous subcorpora, L2 students seem to use a few PVs much more frequently than L1 students do, followed by a rapid drop in frequency for the PVs that follow. Interestingly, it is the PVs *grow up* (NWU and WITS) and *end up* (NWU and UP) for whom this is true, with *end up* and *grow up* having extreme values in the NWU L2 subcorpus, *end up* having an extreme value in the UP L2 subcorpus, and *grow up* having an extreme value in the WITS subcorpus. While these two PVs also appear in the L1 subcorpora (*end up* being in the eleventh position in the WITS L1 subcorpus), they are not used to the same extent as in the L2 subcorpora. The reason for their popularity among L1 students is hard to surmise without further research, such as, for example, the presence of these PVs in the popular media. One known fact about them, as discussed in §3.8.2.5, is that they fall into the category of PVs for which there are no suitable synonyms.

In this section, the PV use of the three institutions represented in the corpora was investigated. L1 and L2 PV use was examined separately, whereafter the two groups were compared to determine the differences in PV use between them. PV use was measured according to the same parameters throughout. The same parameters will be used to compare PV use across the three institutions in the next section.

4.3 Patterns of PV use across three tertiary institutions

In this section, PV use will be compared among the three institutions. Such comparisons are necessary in order to get a more complete picture of student PV use at South African universities, rather than to draw conclusions based on the data from one institution alone. Each institution is subdivided into an L1 and L2 group, and all comparisons between the three institutions will take these subdivisions into account. It should be noted that, while the NWU and UP subcorpora are both derived from the SAMuLCAT corpus, the WITS corpus is a standalone corpus, though also subdivided into L1 and L2 subcorpora. Figure 4.5 provides a visual overview of the structure of this section.

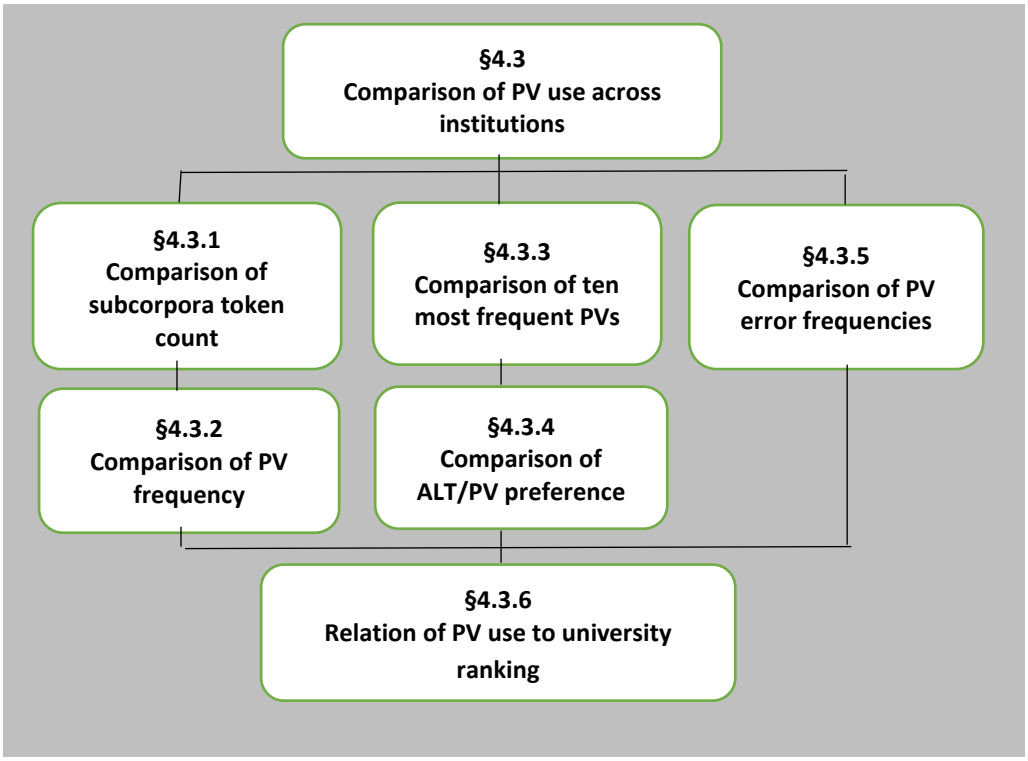

Figure 4.5: Structure of §4.3

To obtain as complete a picture as possible of the patterns of PV use across the three institutions, various comparisons will be made, as was done in the previous section. These are related to the subcorpora token count, PV token and type frequencies, the ten most frequently used PVs, ALT or PV use preference, and PV error frequencies. On the basis of these analyses, the possibility of a link between PV use and university ranking will be investigated.

4.3.1 Comparison of subcorpora sizes across institutions

In Table 4.43, the subcorpora are compared according to size (number of tokens). In total, the corpora of the three institutions comprised 32 101 058 tokens. The NWU subcorpus consists of the highest number of tokens, at 16 929 327. This is largely due to the size of its L2 group, which has 14 898 759 tokens, compared to its L1 group, with only 2 030 578 tokens. The second largest token count is found in the WITS corpus, which has 8 424 209 tokens in total, followed by the UP subcorpus, with 6 747 522 tokens in total.

Table 4.43: Comparison of subcorpus sizes (tokens)

Subcorpora		***L1 / L2 subcorpus size***	***Overall subcorpus size***
NWU	L1	2 030 568	16 929 327
	L2	14 898 759	
UP	L1	2 033 648	6 747 522
	L2	4 713 874	
WITS	L1	5 087 818	8 424 209
	L2	3 336 391	
		Total	**32 101 058**

Figure 4.6 presents a stacked bar chart as a graphical representation of the difference in size of the subcorpora, showing the distribution of the L1 and L2 subcorpora within each institution.

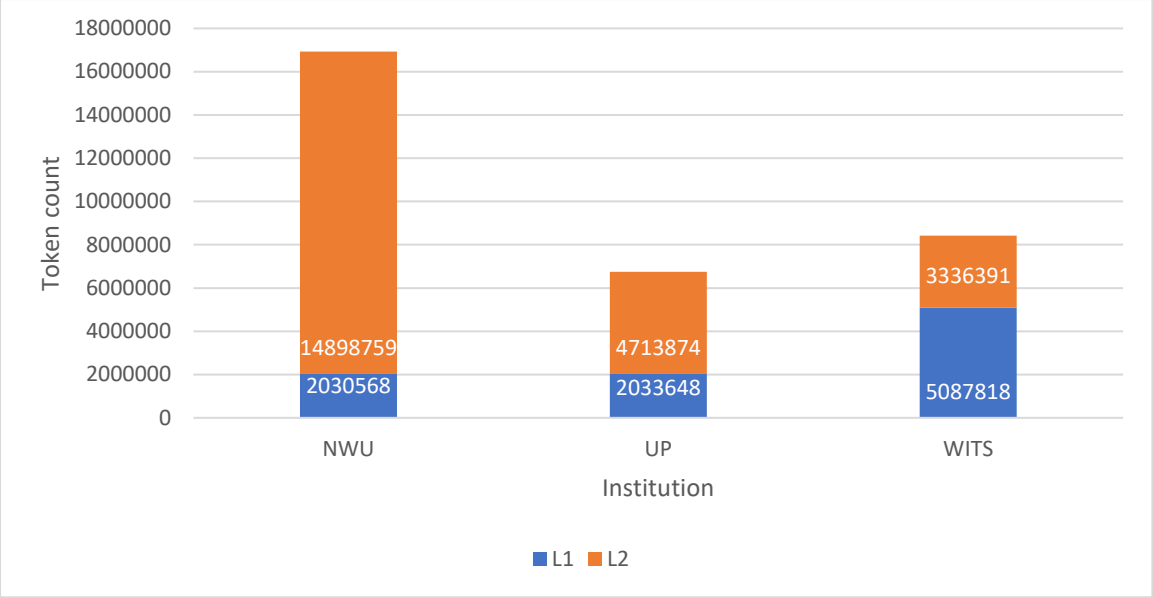

Figure 4.6: Comparison of subcorpora sizes across institutions

The bar chart provides a visual representation of the information previously provided regarding the differing sizes of the subcorpora, showing that NWU has the largest L2 subcorpus, and that WITS has the largest L1 subcorpus.

4.3.2 Comparison of PV use across institutions

PV use across institutions will be compared at different levels to gain as accurate a picture as possible, and, furthermore, to determine whether there is a relationship between overall PV use and university ranking. Firstly, PV representation per institution will be investigated (*absolute PV tokens per subcorpus / tokens per subcorpus * 100*). Table 4.44 shows that the percentage of PVs in the WITS subcorpus, at 0.084%, is the highest of the three institutions, followed by NWU at 0.028%, and UP at 0.023%. Thus, WITS PV use, percentage wise, can be said to be three times higher than that of NWU, and approximately three and a half times higher than that of UP. Of all the students represented, it would seem that WITS students are most likely to use PVs.

Table 4.44: Comparison of PV representation per institution

Subcorpus	***Tokens per subcorpus***	***Absolute PV tokens per subcorpus***	***Percentage of PV tokens to subcorpus tokens***
NWU	16 929 327	4 823	0.028%
UP	6 747 522	1 556	0.023%
WITS	8 424 209	7 053	0.084%

However, it should be noted that PV representation within all three institutions may be considered as low (below 0.1%). To put this into perspective, PV use in the South African corpora was compared to those used by Chen (2013) in her research. Information supplied by Chen (2013:426) indicates that the PV percentage of the LOCNESS-US, comprising essays by American university students, is 0.42%, that of the *General Studies* corpus (GS-UK), comprising essays by British secondary school leavers, is 0.29%, and that of a Chinese student corpus, comprising the argumentative essays of a group of Chinese university students, is 0.30%, all of which are at least 245% (or approximately three and a half times) higher than the WITS corpus PV use. It is also notable that the American and British corpora mentioned here indicate a PV use that is about ten times higher than the PV use of the NWU and UP corpora.

While the LOCNESS-US and GS-UK corpora are native speaker corpora, the Chinese student corpus is not. As the South African corpora include both native and non-native speakers, the comparison is valid, and shows that PV use does not seem to feature to a notable degree in the academic writing of students at the three South African institutions represented. All references to “higher” and “lower” PV use within the South African context should be seen against the backdrop of this comparison with other groups of L1 and L2 students internationally.

Next, relative PV frequency per subcorpus, as well as relative PV frequency per L1 and L2 subcorpora, will be compared (Table 4.45, below). Relative PV frequency per subcorpus is not calculated by the addition of the relative L1 and L2 frequencies, but rather by the addition of the raw (absolute) L1 and L2 frequencies, the total of which is then normalised. The WITS corpus is shown to have the highest

relative PV token frequency (837.23), followed by the NWU subcorpus (284.89), and the UP subcorpus (230.60). Thus , NWU PV use is approximately a third (34.03%) of WITS PV use, and UP PV use if approximately a quarter (27.54%) of WITS PV use. This confirms the impression given in Table 4.44 that, compared to the other students in the study, WITS students are most likely to use PVs.

Table 4.45: Comparison of PV count across institutions

Subcorpus		Relative PV type frequency*	Overall relative PV type frequency#	Relative PV token frequency*	Overall relative PV token frequency#	Percentage of PV types to PV tokens**
NWU	L1	69.93	29.59	268.89	284.89	10.38%
	L2	24.10		287.07		
UP	L1	70.32	50.39	235.05	230.60	21.85%
	L2	41.79		228.69		
WITS	L1	70.95	80.24	773.02	837.23	9.58%
	L2	94.41		935.14		

*Per million words **Of normalised values #Calculated using raw frequency data

Analysing the PV token frequencies of the L1 and L2 subcorpora separately yields further interesting insights. Figure 4.7 provides an illustration of these frequencies. If one were to rank the relative PV token frequencies, the following pattern will be observed: WITS L2, WITS L1, NWU L2, NWU L1, UP L1, UP L2. The UP subcorpus, therefore, is the only subcorpus where the PV L1 relative token frequency is higher than the PV L2 relative token frequency. The difference, which is relatively small (6.36 per million words), can be described in two ways. Firstly, the L1 value (235.05) represents 50.68% of the total, while the L2 value (228.69) represents 49.31% of the total, a 1.37 percentage point difference. Secondly, this can also be expressed as the L1 value being 2.70% higher than the L2 value. Both the other L2 subcorpora show higher relative PV token frequencies than their L1 counterparts. The difference between the NWU L2 and L1 subcorpora (18.18 per million words) can be expressed as a percentage point difference of 3.28, and as a percentage difference of 6.76%. The difference between the WITS L2 and L1 subcorpora (162.12 per million words) equates to a percentage point difference of 9.5, which can also be expressed as a percentage difference of 20.97%. These differences tell us that, of the three corpora, the WITS L2 students are most likely to prefer PVs compared to their L1

counterparts. In the UP subcorpus, the L1 students appear to use more PVs than the L2 students, but by so small a margin that no inference can be drawn. Although the margin between the frequencies of the NWU L1 and L2 subcorpora is larger than that of the UP subcorpus, and suggests that the L1 students in this subcorpus use more PVs than their L2 counterparts, the difference is still relatively small, and assumptions should be treated with caution.

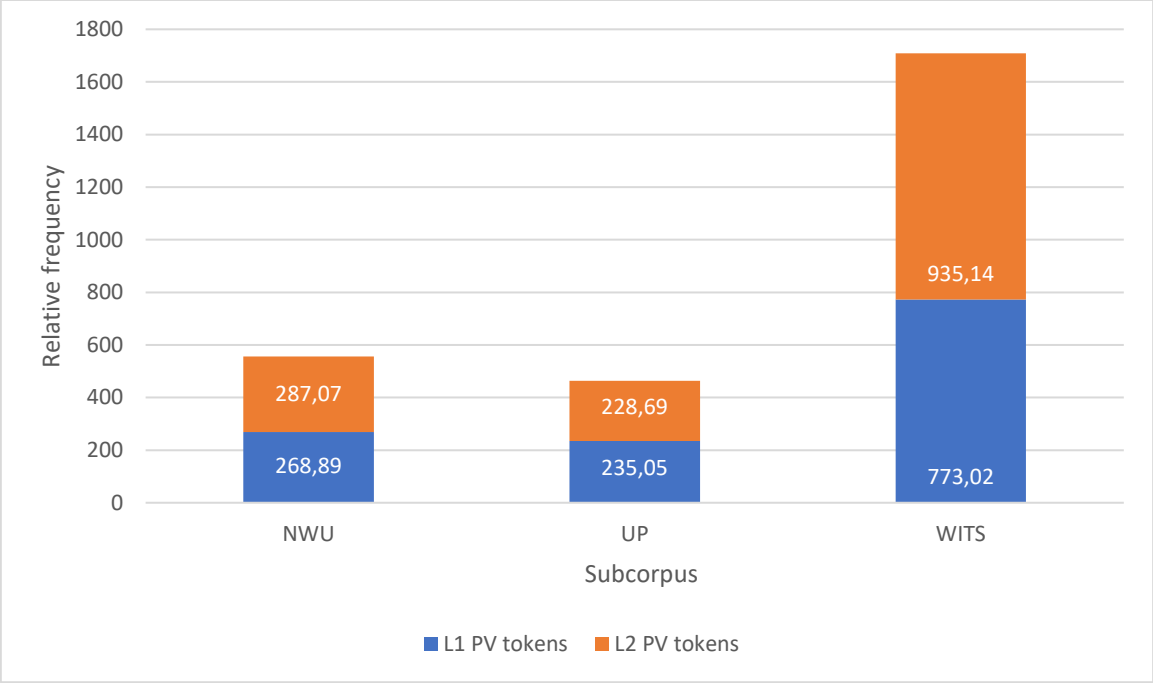

Figure 4.7: Comparison of L1 and L2 relative PV token frequencies

Besides PV token frequency, a comparison of PV type frequency across the subcorpora, as shown in Table 4.45, should also be considered. The WITS corpus has the highest relative PV type frequency (80.24), followed by the UP subcorpus (50.39), and the NWU subcorpus (29.59). However, the relevance of these frequencies will only be evident when considering the percentage of PV types to PV tokens (*relative PV types / relative PV tokens * 100*), which indicates the variety of PVs used. Thus, the WITS subcorpus is shown to have the lowest percentage of PV types to PV tokens (9.58%). In contrast, the percentage of UP PV types to PV tokens is 21.85%, which is the highest of the three institutions, and indicates that the UP students used the greatest variety of PVs, while the WITS

students used the smallest variety of PVs. However, one cannot deduce from this that the UP students are the most adept at PV use, as the number of occurrences of each PV type is not given here. For instance, PV types that occur only once suggest incidental use, rather than established knowledge of the PV.

Investigating PV type frequencies at the L1 and L2 level again reveals interesting results (illustrated in Figure 4.8 below). Two of the L1 corpora have a higher relative incidence of PV type use, namely the NWU L1 subcorpus, with 69.93 PV types compared to the 24.10 PV types of the L2 subcorpus, and the UP L1 subcorpus, with 70.32 PV types compared to the 41.79 PV types of the L2 subcorpus. In these subcorpora, the higher presence of PV types indicates that the L1 students use a greater variety of PVs than their L2 counterparts. In contrast, in the WITS corpus, the L2 subcorpus has a higher relative PV type frequency (94.41) than the L1 subcorpus (70.95). In this case, the L2 WITS students seem to have a higher variety of PV use. Nevertheless, as discussed in the previous paragraph, this does not necessarily provide evidence of fluency in PV use. The graph below (Figure 4.9) depicts the difference in PV type use between the L1 and L2 subcorpora.

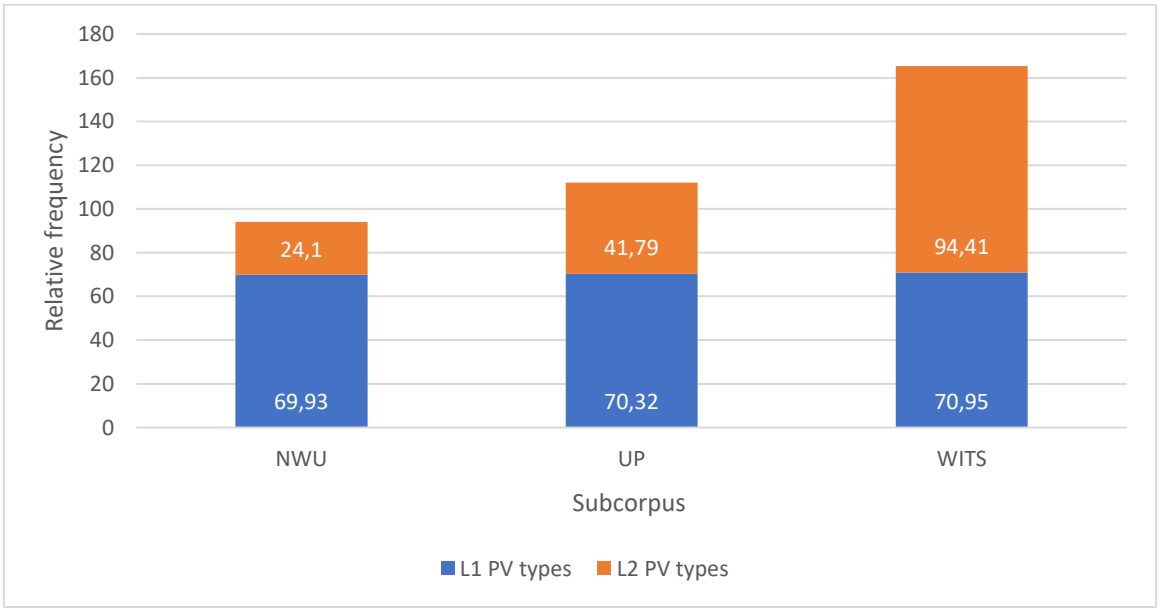

Figure 4.8: Comparison of L1 and L2 relative PV type frequencies

PV tokens and types used in the three institutions can be summarised as follows. The WITS corpus has the highest PV token use, but the lowest PV type frequency, suggesting that WITS students use the smallest variety of PVs. On the other hand, the UP corpus, which has the lowest PV token frequency, has the highest PV type frequency, suggesting that, though the students do not use as many PVs as the students at the other institutions, they use the greatest variety of PVs. However, PV variety does not necessarily equate to PV fluency, as this can only be established by investigating the occurrence of specific PVs across institutions (see §4.3.3).

4.3.3 Comparison of ten most frequently used PVs across institutions

Table 4.46 gives a comparison of the ten most frequently used PVs across the three subcorpora, broken down per L1 and L2 groups. This gives an indication of PVs that are in general use among students across these institutions, compared to PVs that occur at a single university only. For example, the PV *end up* is the most used PV for two of the institutions represented (NWU and UP), in both their L1 and L2 groups, but only features in the L2 group of the WITS corpus (4th position). The PV *carry out* is found in all the subcorpora, for both L1 and L2 groups: WITS L1 (1st position), NWU L1 (2nd position), WITS L2 (2nd position), NWU L2 (3rd position), UP L2 (5th position), and UP L1 (6th position). On the other hand, the PV *act out* appears once only, in the WITS L2 subcorpus (4th position).

Table 4.46: Comparison of ten most frequently used PVs across all three subcorpora

NWU						***UP***						***WITS***					
L1			***L2***			***L1***			***L2***			***L1***			***L2***		
PV	***Absolute frequency***	***Relative frequency****	***PV***	***Absolute frequency***	***Relative frequency****	***PV***	***Absolute frequency***	***Relative frequency****	***PV***	***Absolute frequency***	***Relative frequency****	***PV***	***Absolute frequency***	***Relative frequency****	***PV***	***Absolute frequency***	***Relative frequency****
END UP	29	14.28	END UP	582	39.06	END UP	32	15.74	END UP	120	25.46	CARRY OUT	325	63.88	GROW UP	411	123.19
CARRY OUT	28	13.79	GROW UP	339	22.75	BRING ABOUT	20	9.83	GROW UP	77	16.33	MAKE UP	303	59.55	CARRY OUT	217	65.04
BRING ABOUT	25	12.31	CARRY OUT	160	10.74	MAKE UP	20	9.83	POINT OUT	49	10.39	GROW UP	295	57.98	MAKE UP	201	60.24
COME UP	24	11.82	FIND OUT	154	10.34	POINT OUT	19	9.34	MAKE UP	32	6.79	ACT OUT	138	27.12	END UP	160	47.96

TAKE ON	OPEN UP	SET OUT	GO ON	GROW UP	MAKE UP
14	15	18	18	22	24
7.00	7.39	8.86	8.86	10.83	11.82
SEND OUT	GIVE OUT	STAND OUT	SET OUT	GO ON	BRING ABOUT
72	97	98	100	119	134
4.83	6.51	6.58	6.71	7.99	8.99
GO ON	BUILD UP	FIND OUT	SLOW DOWN	CARRY OUT	GROW UP
11	11	15	16	17	18
5.41	5.41	7.38	7.87	8.36	8.85
BRING UP	BREAK DOWN	GO ON	TAKE OVER	BRING ABOUT	CARRY OUT
19	22	24	26	28	31
4.03	4.67	5.09	5.52	5.94	6.58
POINT OUT	SET OUT	OPEN UP	FIND OUT	PICK UP	BRING UP
77	80	80	108	113	133
15.13	15.72	15.72	21.23	22.21	26.14
TURN OUT	OPEN UP	COME UP	POINT OUT	PICK UP	FIND OUT
48	48	52	57	75	124
14.39	14.39	15.59	17.08	22.48	37.17

Similarly coloured blocks show presence of specific top ten PVs across corpora; white blocks show top ten PVs that appear only once across corpora. *Per million words

Table 4.47 gives a summary of the information about the most frequently used PVs across the subcorpora (provided in Table 4.46 above). Besides the relative frequency, information is also provided as to the subcorpora in which a PV is found, and the *range %* of each PV (discussed below). PVs are ranked according to relative frequency, rather than number of occurrences, as the ranking of these two categories does not necessarily correspond. For instance, the PV *pick up* appears in fewer subcorpora (two) than the PV *bring about* (four), yet *pick up* has a higher relative frequency (44.69) than *bring about* (37.07).

Table 4.47: Summary of most frequently used PVs across subcorpora, with range percentage

PV	***Relative frequency****	***Number of subcorpora in which PV occurs***	***NWU***		***UP***		***WITS***		***Range %*****
			L1	***L2***	***L1***	***L2***	***L1***	***L2***	
GROW UP	239.93	6	•	•	•	•	•	•	**100**
CARRY OUT	168.39	6	•	•	•	•	•	•	**100**
MAKE UP	148.23	5	•		•	•	•	•	**83**
END UP	142.50	5	•	•	•	•		•	**83**
FIND OUT	76.12	4		•	•		•	•	**67**
POINT OUT	51.94	4			•	•	•	•	**67**
PICK UP	44.69	2					•	•	**33**
OPEN UP	37.50	3	•				•	•	**50**
BRING ABOUT	37.07	4	•	•	•	•			**67**
SET OUT	31.29	3	•	•			•		**50**
BRING UP	30.17	2				•	•		**33**
COME UP	27.41	2	•					•	**33**
GO ON	27.35	4	•	•	•	•			**67**
Subtotal of PVs with multiple occurrences across subcorpora			**9**	**7**	**8**	**8**	**9**	**9**	**–**
ACT OUT	27.12	1					•		**17**
TURN OUT	14.39	1						•	**17**
SLOW DOWN	7.87	1			•				**17**
TAKE ON	7.00	1	•						**17**
STAND OUT	6.58	1		•					**17**
GIVE OUT	6.51	1		•					**17**
TAKE OVER	5.52	1				•			**17**
BUILD UP	5.41	1			•				**17**
SEND OUT	4.83	1		•					**17**
BREAK DOWN	4.67	1				•			**17**
Subtotal of PVs with single use across subcorpora			**1**	**3**	**2**	**2**	**1**	**1**	**_**
Total, confirming top ten PVs of each corpus are shown			**10**	**10**	**10**	**10**	**10**	**10**	**_**

*Per million words ** Range % is used as a measure of dispersion, to show the percentage of corpora in which the PV occurs

The table is divided into two parts to differentiate between PVs that appear twice or more (of which there are 13) in the top ten PVs across the three subcorpora, and those that appear once only (of which there are ten). This information is important as "the more frequent and evenly distributed the word is, the more prominent it is considered to be" (Brezina, 2018:54). The PVs *grow up* and *carry out* appear in all six subcorpora, which suggests that these PVs are in general use among students (both L1 and L2) across institutions. The PVs *make up* and *end up* appear in five of the six subcorpora and can, thus, be considered to be in general use. Although it is conspicuous that *make up* does not appear in the NWU L2 subcorpus (the largest subcorpus), and that *end up* is absent from the WITS L1 subcorpus (the second largest subcorpus), it should be remembered that only the ten most frequently used PVs in each subcorpus are shown here. Most of the PVs that are absent from the top tier of the table are, in fact, present in the subcorpora from which they appear to be missing, albeit at a lower frequency. The only PVs that are entirely absent from a subcorpus are *pick up* (NWU L1), *bring about* (WITS L2), *bring up* (NWU L1), *come up* (UP L1), and *go on* (WITS L1). This suggests familiarity with high-frequency PVs across the three institutions.

In contrast to the PVs that appear across several subcorpora, the PVs *pick up*, *bring up* and *come up* are indicated as being present in only two subcorpora. Furthermore, in the case of the PV *pick up*, the two subcorpora are from the same institution (NWU L1 and L2). Likewise, some of the PVs that are present in four subcorpora (*point out*, *bring about*, and *go on*), are in effect only present in two institutions (UP and WITS, NWU and UP, and NWU and UP, respectively). If PV presence is confined to only one or two institutions, then the PV cannot be said to be widely used across South African universities. Within this list of the most frequently used PVs across the subcorpora, there are ten that appear in only a single subcorpus (see Table 4.47 above).

Frequency reporting should be accompanied by information on dispersion (Brezina, 2018:53). A simple measure of dispersion that can be used is *range %*. This indicates the spread of PVs across the subcorpora, given as a percentage, calculated as

*(number of occurrences per PV across subcorpora / number of subcorpora) * 100*

Using the PVs *pick up* and *bring about* as examples (Table 4.47 above), we see that, although *pick up* has a higher relative frequency, its *range %* is only 33, compared to *bring about*, with a *range %* of 67. Thus, *bring about* is more widely dispersed among the subcorpora than is *pick up*. However, the *range %* measurement provides only basic dispersion information about a PV, referring simply to the presence or absence of the PV in a corpus, and not actual frequencies (Brezina, 2018:48). Therefore, while this shows whether PVs are used across institutions, it does not indicate to what extent such PVs are used.

It is, consequently, necessary to include other descriptive statistics, such as measures of central tendency (see Table 4.48 below). This refers to the mean, median and mode, which show the central tendency of the data. The *Descriptive Statistics Calculator* (Furey, 2024) was used to calculate the mean and median of each of the top ten PVs. (The mode, being the value that occurs most often in the data set, was not deemed necessary for inclusion in the table, as it did not add value to the discussion in this particular instance. This is because the values are non-discrete, and there are no duplicates in any of the datasets, resulting in every value being the mode. Therefore, no useful information about the dataset is provided by the mode.) The mean (*sum of values / number of subcorpora*) is not always a reliable measurement, because "in distributions with outliers, the mean might represent the outlier more than the rest of the values" (Brezina, 2018:10). Therefore, extreme values distort the mean. In such cases, the median (*middle value of series of values ranked from lowest to highest*) is a better measure, because it is always in the middle of the distribution.

When discussing dispersion, it is also necessary to include standard deviation, as it indicates how actual frequencies are dispersed from the mean. Sample standard deviation is used here, as the data represent a sample of the population rather than the whole population. As with the mean and median, the *Descriptive Statistics Calculator* (Furey, 2024) was used to calculate the sample standard deviation, based on the following formula:

$$\sqrt{\frac{\textit{sum of the squared differences between each number and the mean}}{\textit{number of subcorpora - 1}}}$$

The sample standard deviation for each of the top ten PVs across the three subcorpora under discussion is given in Table 4.48 below. This statistical measurement was not included for PVs represented in one subcorpus only, as such data would inevitably be skewed.

Calculating skewness will indicate the skewness of the data set, showing distortion of the mean because of extreme values. The formula for this is *(3*(mean – median) / standard deviation (SD))*. A positive result shows that the data are positively (right) skewed, and a negative result shows that the data are negatively (left) skewed. For the moment, this is of less importance than the resulting value: a result equal to or less than 0.5 indicates that the data are not skewed, a result between 0.5 and 1 shows that the data are slightly skewed, and a result of more than one indicates that the data are highly skewed. The skewness column in Table 4.48 below gives the results of this calculation, and shows that the data are symmetrical in only one instance (indicated in green), that there are five instances where the data are slightly skewed (indicated in orange), and that the data are highly skewed in seven instances (indicated in yellow). The highly skewed data linked to certain PVs (*grow up*, *carry out*, *make up*, *pick up*, *open up*, *bring up*, *come up*) result from the high PV frequencies of these PVs in the WITS subcorpora, indicating that high-frequency PVs are most often found at this institution. This confirms the high PV use noted for WITS in §4.3.2. (See §4.3.6 for a discussion of the relationship between ranking and PV use.)

Table 4.48: Dispersion of most frequently used PVs across subcorpora

PV	NWU		UP		WITS		Mean	Median	Sample SD	Skewness
	L1	L2	L1	L2	L1	L2				
GROW UP	10.83	22.75	8.85	16.33	57.98	123.19	39.9	19.5	44.5	1.38
CARRY OUT	13.79	10.74	8.36	6.58	63.88	65.04	28.0	12.2	28.3	1.67
MAKE UP	11.82	0.00	9.83	6.79	59.55	60.24	24.7	10.8	27.5	1.52
END UP	14.28	39.06	15.74	25.46	0.00	47.96	23.7	20.6	17.5	0.53
FIND OUT	0.00	10.34	7.38	0.00	21.23	37.17	12.6	8.8	14.3	0.80
POINT OUT	0.00	0.00	9.34	10.39	15.13	17.08	8.6	9.8	7.2	-0.50
PICK UP	0.00	0.00	0.00	0.00	22.21	22.48	7.4	0	11.5	1.93
OPEN UP	7.39	0.00	0.00	0.00	15.72	14.39	6.2	3.6	7.4	1.05
BRING ABOUT	12.31	8.99	9.83	5.94	0.00	0.00	6.1	7.4	5.2	-0.75
SET OUT	8.86	6.71	0.00	0.00	15.72	0.00	6.1	7.4	5.2	-0.75
BRING UP	0.00	0.00	0.00	4.03	26.14	0.00	5.0	0	10.4	1.44
COME UP	11.82	0.00	0.00	0.00	0.00	15.59	4.5	0	7.1	1.90
GO ON	8.86	7.99	5.41	5.09	0.00	0.00	4.5	5.2	3.8	-0.55
ACT OUT	0.00	0.00	0.00	0.00	27.12	0.00	4.5	0	11.0	
TURN OUT	0.00	0.00	0.00	0.00	0.00	14.39	2.3	0	5.8	
SLOW DOWN	0.00	0.00	7.87	0.00	0.00	0.00	1.3	0	3.2	
TAKE ON	7.00	0.00	0.00	0.00	0.00	0.00	1.3	0	3.2	
STAND OUT	0.00	6.58	0.00	0.00	0.00	0.00	1.0	0	2.6	
GIVE OUT	0.00	6.51	0.00	0.00	0.00	0.00	1.0	0	2.6	
TAKE OVER	0.00	0.00	0.00	5.52	0.00	0.00	0.9	0	2.2	
BUILD UP	0.00	0.00	5.41	0.00	0.00	0.00	0.9	0	2.2	
SEND OUT	0.00	4.83	0.00	0.00	0.00	0.00	0.8	0	1.9	
BREAK DOWN	0.00	0.00	0.00	4.67	0.00	0.00	0.7	0	1.9	
Total	**106.96**	**124.5**	**88.02**	**90.8**	**324.68**	**417.53**				

As the table makes information available about individual PV frequencies across institutions, it is possible to consider the relevance of PV types to PV tokens (as discussed in §4.3.2) with more accuracy. The UP subcorpus was indicated as having the highest PV type percentage to PV tokens

(24.18%). This showed that the UP students used the greatest variety of PVs, but did not offer proof of the UP students' actual knowledge of PVs. The totals per subcorpus in Table 4.48 show that the UP L1 and L2 subcorpora have the lowest relative frequencies for the top ten PVs (88.02 and 90.8, respectively) of all the subcorpora. This suggests that, though the UP students use a variety of PVs, they do not use them on a regular basis. On the other hand, it would appear that the WITS students, though having the lowest percentage of PV types to PV tokens of all the subcorpora, but the highest relative frequencies for the top ten PVs, use specific PVs on a regular basis, suggesting established knowledge of the PVs used. Of course, these totals reflect only the totals for the top ten PVs of the institutions. However, it is possible to draw reasonably reliable conclusions based on this sample, as the remaining PVs in the subcorpora have progressively lower frequencies.

Finally, it is interesting that Zipf's law of rapidly diminishing word frequency (described in §4.2.1.1.2) is only evident in the L2 subcorpora, a nd not the L1 subcorpora. This suggests that L2 PV use is more dispersed that that of L1 PV use. This tendency was also observed in the box and whisker plots, where the L2 data consistently displayed wider dispersion than the L1 data, which tended to display limited dispersion. The L2 data also had more dispersed extreme values, showing that the high scores were scattered, in contrast to the L1 data, where this was not a feature. On the other hand, for both L1 and L2 data, the extreme values at the lower level were mostly aligned with the lower quartile, suggesting that these values were not widely dispersed. This indicates that frequency of use of the PVs at the bottom of the table was much closer than those PVs at the top of the table, as observed in the case of both L1 and L2 PV frequencies.

L2 students across the institutions, therefore, seem to use the highest frequency PVs often, but then show a sharp decline in PV use. This is, for example, illustrated for the PV *grow up* in Table 4.48. This PV was shown not to have suitable ALTs, which might, in part, explain its presence at the top of the table, but does not address its high use among L2 students.

A final note in this section refers to PVs found in the subcorpora that could be considered slang, thus, PVs that should be used only in informal registers (speech and informal writing). Examples of some of these PVs are *bandy around*, *beef up*, *boss around*, *churn out*, *doze off*, *fork out*, *gobble up*, *man up*, *mess around*, *plunk down*, *rev up*, *slack off*, *suss out*, and *zone out*. The frequencies of these PVs were low (not more than one occurrence per PV per subcorpus), and, moreover, these PVs were not present in all the subcorpora. Thus, presumably, they are not a cause for concern. Nevertheless, their presence in student academic writing reveals that students are unaware of the fact that some PVs should not be used in an academic register.

4.3.4 Comparison of PV and ALT use across institutions

Table 4.49 compares the relative values of the top ten PVs to their ALTs across the three subcorpora used in this study. The total relative PV and ALT frequencies per subcorpus are calculated using absolute values, which are then normalised. These total values, therefore, are not derived from totalling the L1 and L2 values.

The following observations can be made. In each of the three subcorpora, the ALT frequency is higher than that of the PV frequency. The WITS corpus has the highest number of ALTs per million words, followed by the UP corpus, and, lastly, the NWU corpus. This suggests that, of the three institutions, the WITS students represented in the corpus most prefer ALTs to PVs. However, a different picture emerges when the ratios of PV to ALT of the different institutions are considered. The UP ratio of 1:7.8 PVs to ALTs is the highest, followed by WITS, with a ratio of 1:5.1. NWU has the lowest ratio of 1:2.7. This suggests that, contrary to the impression created by looking at the ALT frequency only, students represented in the UP subcorpus are most likely to use ALTs rather than PVs. This is confirmed when considering the ratios of the PVs to ALTs of the L1 and L2 subcorpora for each institution, since the UP subcorpora once again have the highest ratios of all three institutions. This preference for ALTs by UP students seems to align with UP having the lowest PV token frequency of the three institutions (§4.3.2).

Table 4.49: Comparison of relative frequencies of top ten PVs and ALTs

NWU				***UP***				***WITS***			
PV frequency*		***ALT frequency****		***PV frequency****		***Alt frequency****		***PV frequency****		***Alt frequency****	
L1	***L2***	***L1***	***L2***	***L1***	***L2***	***L1***	***L2***	***L1***	***L2***	***L1***	***L2***
106.87	124.51	617.07	291.84	88.02	90.80	935.76	602.90	324.68	417.53	1937.20	1681.17
122.39		330.85		89.96		703.22		361.46		1 835.78	
Ratio											
1:2.7				1:7.8				1:5.1			
Ratio L1 PV to L1 ALT											
1:5.8				1:10.6				1:5.9			
Ratio L2 PV to L2 ALT											
1:2.3				1:6.6				1:4.0			

*Per million words

Notably, the L1 subcorpora of all three institutions have higher PV to ALT ratios, which suggests that L1 students are more likely than L2 students to prefer the use of ALTs. Thus, the lower PV token frequency of the three L1 subcorpora compared to the L2 subcorpora suggests that L1 students prefer a more formal academic writing style, as it is unlikely that this can be ascribed to PV avoidance behaviour.

In §4.3.5, PV error frequencies will be compared across the three institutions.

4.3.5 Comparison of PV error frequencies across institutions

Table 4.50 compares L1 and L2 error frequency, as well as total error frequency, across the three subcorpora. As discussed in §4.3.2, relative PV frequency per subcorpus is calculated using raw or absolute values, rather than totalling the relative L1 and L2 frequencies. The highest error frequency is found in the WITS corpus, at 32.41 occurrences per million words. This suggests that students at this institution were more inclined to make mistakes in the use of PVs than students at the other institutions, with a relative frequency of 16.59 for the UP subcorpus, and a relative frequency of 14.65 for the NWU subcorpus. However, for a more dependable understanding of what the error frequency

represents, it needs to be considered in the context of the PV tokens of the relevant institutions. Because of the PV error frequency not being included in the PV frequency (the PV frequency needed to reflect correctly used PVs), it needs to be added before calculating the percentage of PV errors to PV tokens. Therefore, the formula reads as follows: *Relative subcorpus error frequency / (relative subcorpus PV frequency + relative subcorpus error frequency) * 100.* From this calculation it emerges that the highest percentage of PV errors to PV tokens is, in fact, present in the UP subcorpus.

Table 4.50: Comparison of subcorpora error token frequency

Subcorpus		***Relative error frequency****	***Relative subcorpus error frequency#***	***PV errors as percentage of total PV tokens****
NWU	L1	4.92	14.65	4.89%
	L2	15.97		
UP	L1	9.83	16.59	6.71%
	L2	19.52		
WITS	L1	28.89	32.41	3.73%
	L2	37.78		

*Per million words **Of normalised values #Calculated using raw frequency data **In this case, total PV tokens=PV tokens + PV errors

Further information on PV error distribution is provided by the L1 and L2 groups. Differences in error distribution between the L1 and L2 groups of each subcorpus may be interpreted by means of the relative error frequency. For instance, the error frequency of the NWU L2 group is 225% higher than that of the NWU L1 group (15.97 / 4.92 * 100) - 100), that of the UP L2 group is 99% higher than that of the UP L1 group (19.52 / 9.83 * 100) - 100), and that of the WITS L2 group is 31% higher than that of the WITS L1 group (37.78 / 28.89 * 100) - 100). Therefore, it seems that the NWU L1 and L2 students have the largest variance in PV proficiency, and that the WITS students have the lowest, suggesting that the WITS L1 and L2 students demonstrate similar proficiency levels.

4.3.6 Comparison of subcorpus PV use to university ranking

The *Times Higher Education Sub-Saharan Africa University Rankings* (2024) uses teaching, research environment, research quality, industry, and international outlook as indicators in the ranking of university performance. At the start of this research study, it had been considered viable to correlate

PV use with university ranking. This was not due to the expectation that students would have received instruction regarding PV use at a higher ranked university, as, first of all, a large proportion of student writing incorporated into the corpora are of first-year students, and insufficient time will have elapsed for such teaching (if any) to have had an impact. Secondly, there is not sufficient clarity about the stance of educational authorities in South Africa towards PV use in academic writing to draw conclusions as whether students at higher ranking universities would be expected to use more or fewer PVs than lower ranked universities. Rather, PV fluency as an indicator of competent English use was expected to be observed at higher ranked universities, since students who had performed well at school (to which English proficiency is often linked) would be expected to be attracted to studying at higher ranking universities.

In practice, rating universities according to PV use proved to be less straightforward than anticipated, because of the number of variables that have an impact on results. PV use was measured according to five categories: PV token frequency, PV type frequency, frequency of most used PVs, PV to ALT frequency, and PV error frequency. By taking overall performance into consideration, the WITS students can be said to be the most highly rated as far as PV use is concerned, as a result of having the highest frequencies in two important categories: their percentage of PV use was the highest, and the PVs that they used most often were used at higher frequencies than at NWU and UP. Distinguishing between the other two universities delivers interesting results. UP had the lowest percentage of PV use, as well as the lowest frequency for the PVs used most often compared to the other universities. It also had the highest ratio of ALTs to PVs, which could, on the one hand, indicate PV avoidance, or, on the other hand, a preference for more formal academic writing. However, the high PV error frequency displayed by UP students compared to the NWU and WITS students suggests that high ALT and low PV frequency is the result of poor PV knowledge, rather than of good academic writing. It would seem, therefore, that PV use at UP is the weakest of the three institutions.

Having considered all aspects of PV use at the different universities, it is now possible to assess whether their PV use is aligned with their institutional ranking. Firstly, the level of PV use demonstrated by the WITS students seems to coincide with the ranking of their institution (i.e. the highest ranking of the three institutions represented), as provided by the *Times Higher Education Sub-Saharan Africa University Rankings* (2024) (see Table 4.51 below), when compared to that of UP and NWU students. On the other hand, the PV use demonstrated at UP (ranked second highest of the three institutions represented) does not seem to align with its ranking, as it displayed evidence of having the lowest PV proficiency of the three institutions. It would, therefore, seem that the level of PV proficiency at a university cannot be said to necessarily coincide with its ranking.

*Table 4.51: South African university rankings, compared to all African universities**

Africa rank	*World university rank 2024*	*World university rank 2023*	*University*
2	*302 - 350*	*251 - 300*	*University of Witwatersrand*
5	*501 - 600*	*801 - 1000*	*University of Pretoria*
7	*601 - 800*	*601 - 800*	*North-West University*

* Times Higher Education (2024)

An alternative way of approaching the rankings is to consider how the difference in PV use of L1 and L2 students across universities align with their ranking. Table 4.52 presents the L1 and L2 PV frequencies for the three universities, ranked from highest to lowest. (As it happens, the L1 and L2 frequencies of each of the universities had the same sequence of highest to lowest.) Comparing this information to the rankings in Table 4.51 shows that there is correlation only in one instance. The WITS subcorpus is ranked first of the three universities represented, and has the highest PV frequency in both the L1 and L2 subcorpora of the universities. However, the other two universities do not align with their rankings, as NWU has the second highest PV frequencies for both the L1 and L2 subcorpora, yet is ranked third of the three universities. Likewise, UP is ranked second of the universities, but has the lowest frequencies. This confirms the conclusion drawn in the previous paragraph about PV use and university ranking.

Table 4.52: Comparison of PV token count across institutions, ranked by frequency

Subcorpus	***Relative PV token frequency***	
	L1	***L2***
WITS	773.02	935.14
NWU	268.89	287.07
UP	235.05	228.69

After an examination of PV use across institutions, overall L1 and L2 PV use will now be considered.

4.4 L1 and L2 patterns of PV use

This section presents the culmination of the analyses of the previous sections, in that overall L1 and L2 PV use can now be examined. The L1 and L2 data of the three institutions are collated and compared in order to determine whether the PV profile, as established by the parameters of token frequency, most used PVs, PV / ALT preference, and error frequency, will be different for L1 and L2 students. The findings are then be summarised and discussed in §4.5.

Figure 4.9 presents a visual representation of the analyses to be discussed in this section, and indicates the subsections in which the various aspects of PV use will be examined.

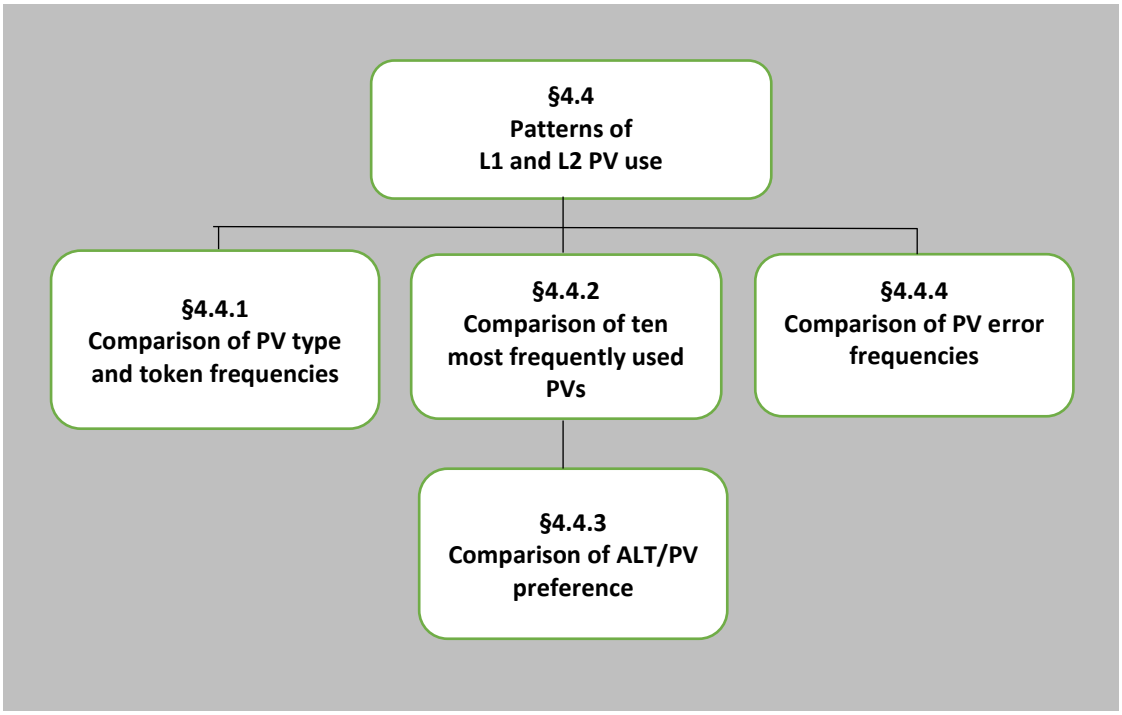

Figure 4.9: Structure of §4.4.

4.4.1 Comparison of L1 and L2 PV type and token frequencies

A comparison of overall PV use by L1 and L2 students is given in Table 4.53. The table provides information about the PV type and token frequencies, as well as the percentage of PV types represented by the PV tokens. An examination of L1 and L2 PV token frequencies provides the following results. The relative PV token frequency for the L2 group (1 450.90) is higher than that of the L1 group (1 276.96) by 13.62% ((1 450.90 – 1 276.96) / 1 276.96 * 100). This fact suggests that L2 students were more likely to use PVs than their L1 counterparts.

Table 4.53: Comparison of L1 and L2 PV type and token frequencies

Group	***Subcorpus token frequency***	***Absolute PV type frequency***	***Absolute PV token frequency***	***Relative PV type frequency****	***Relative PV token frequency****	***Relative PV types as % of relative PV tokens***
L1	9 152 034	646	4 957	211.20	1 276.96	16.54%
L2	22 949 024	871	8 475	160.30	1 450.90	11.05%

*Per million words *Of normalised values

The PV type frequency of the two groups should also be taken into consideration. The PV types reflect the variety of PVs represented within the PV token total. In this case, the PV type frequency of the L1 group is 16.54% of the L1 PV token frequency (211.20 / 1 276.96 * 100), and that of the L2 group is 11.05% of the L2 PV token frequency (160.30 / 1 450.90 * 100). This suggests that the L1 students use a greater variety of PVs (by 5.49 percentage points) than do the L2 students. This information can also be interpreted as a percentage difference, namely that the variety of PVs used by L1 students was 49.68% more than for L2 students (16.54 / 11.05 * 100) - 100). This indicates the substantial difference in PV variety between the two groups.

It might be surmised that this indicates greater PV proficiency by L1 students, especially if using a greater variety of PVs coincides with making fewer errors, as this suggests a better overall understanding of PVs (both in structure and meaning). However, no conclusions can be drawn from the limited information provided here. Rather, this will be addressed when the frequencies of the top

ten most used PVs in the two subcorpora are evaluated (§4.3.3). Figure 4.10 illustrates the difference in PV type and token frequency within the L1 and L2 subcorpora.

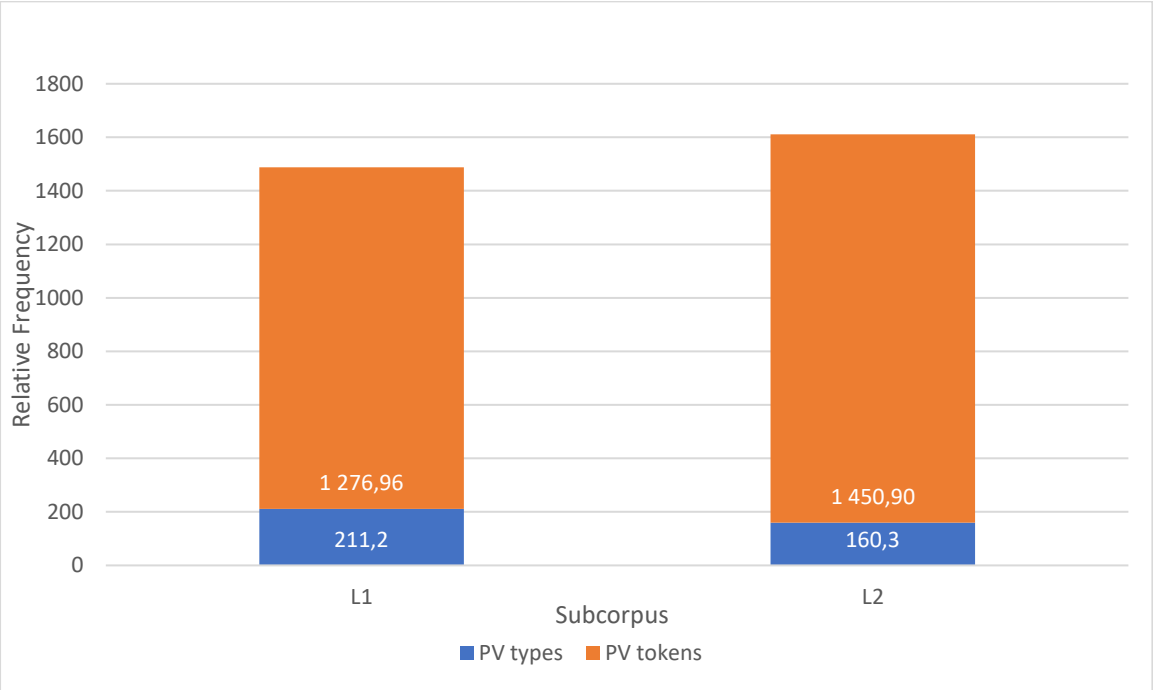

Figure 4.10: Comparison of PV types and tokens

In summary, after examining the PV token data that had been extracted regarding PV use by L1 students, it was observed that the L2 students were 13.62% more likely to use PVs in their writing than the L1 students. This would suggest that there is no notable difference in the use of PVs by L2 students as the results reflect similar levels of proficiency. On the other hand, an investigation of the PV type data indicated that the L1 students were more likely to use particular PVs more than once, suggesting greater familiarity with the use of PVs. Therefore, various factors appear to be at play in an assessment of PV use by a specific group, as will be discussed in §4.5.

4.4.2 Comparison of L1 and L2 ten most frequently used PVs

This subsection compares the L1 and L2 ten most frequently used PVs (see Table 4.54). Keeping in mind that the data presented here are drawn from three different institutions, the results are intriguing in that five PVs (*carry out*, *make up*, *grow up*, *end up*, and *find out*) are the most frequently

used PVs for both groups (although not in the same order, or at the same frequencies). A further two PVs also feature in the top ten PVs of both groups, namely *point out* and *open up*. Such overlap suggests that there are PVs that are in general circulation among students. Of these PVs, there are three that have higher frequencies in the L1 subcorpus (*carry out*, *make up*, and *open up*), and four that have higher frequencies in the L2 subcorpus (*grow up*, *end up*, *find out*, and *point out*). The total frequency of the L2 ten most frequent PVs is higher than the total frequency of the L1 ten most frequent PVs. At face value, this seems to support the results in §4.3.2, where the L2 subcorpus was found to have a higher overall PV frequency than the L1 subcorpus.

Table 4.54: Comparison of L1 and L2 ten most frequently used PVs

L1			***L2***		
PV	***Absolute frequency***	***Relative frequency****	***PV***	***Absolute frequency***	***Relative frequency****
CARRY OUT	370	86.03	GROW UP	827	162.27
MAKE UP	347	81.20	END UP	862	112.48
GROW UP	335	77.66	CARRY OUT	408	82.36
END UP	61	30.02	MAKE UP	233	67.03
FIND OUT	123	28.61	FIND OUT	278	47.51
ACT OUT	138	27.12	POINT OUT	106	27.47
BRING UP	133	26.14	PICK UP	75	22.48
SET OUT	98	24.58	COME UP	52	15.59
POINT OUT	96	24.47	BRING ABOUT	162	14.93
OPEN UP	95	23.11	OPEN UP	48	14.39
Total	**1 796**	**428.94**	***Total***	**3 051**	**566.51**

*Per million words

Figure 4.11 illustrates the distribution of the most frequent PVs across the L1 and L2 subcorpora by means of a box-and-whisker plot. Neither of the subcorpora shows normal distribution since the mean and median values do not align. Furthermore, the data appear positively skewed as most of the values are grouped in the lower quintile. Therefore, neither of the two subcorpora are evenly distributed. However, there are also differences between L1 and L2 in the dispersion of PV frequencies. The interquartile range of the L2 subcorpus is longer than that of the L1 subcorpus, suggesting that the L2

values are more dispersed than the L1 values. Furthermore, the high scores indicated by the whiskers shows that the L2 values are more scattered than the L1 values in the upper range. Therefore, the L2 data are more widely distributed than that of the L1. This indicates a higher frequency of those PVs that are most often used by the L2 students, compared to the L1 students.

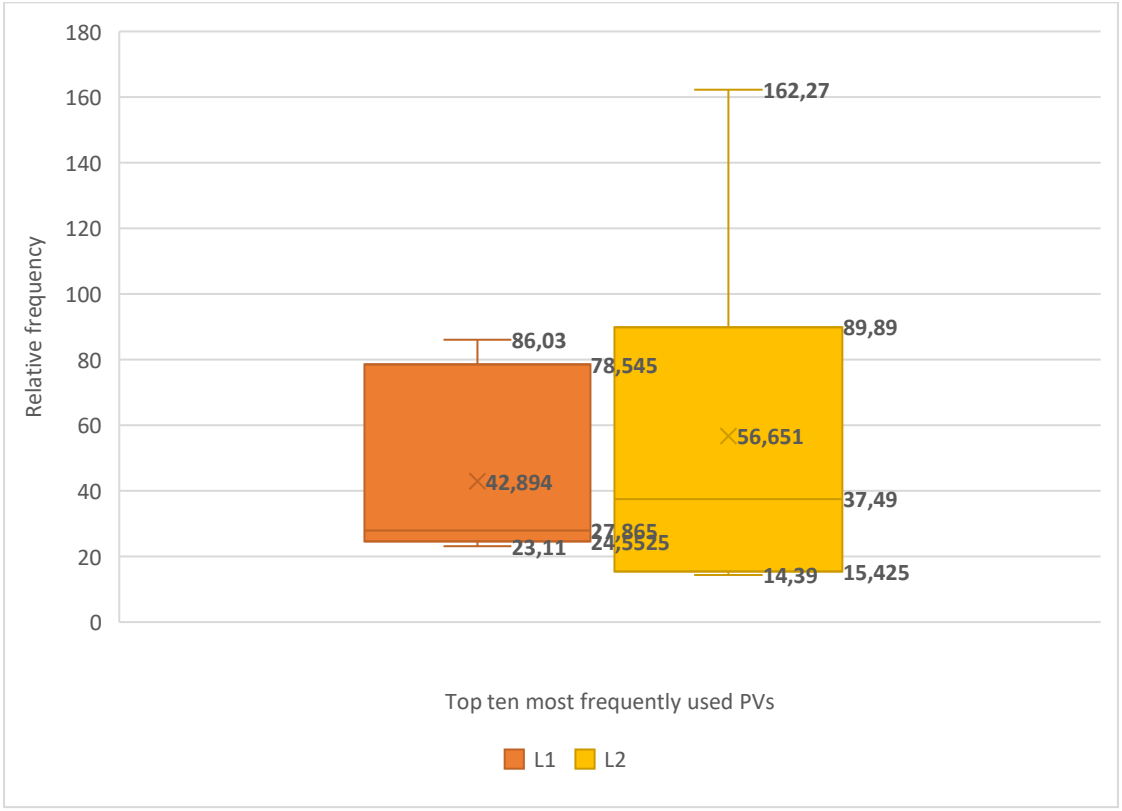

Figure 4.11: Distribution of L1 and L2 top ten PVs

This seems to support the higher relative PV total for the L2 subcorpus compared to the L1 subcorpus, as observed in the table. Nevertheless, the extreme values of the L2 data, illustrated by the box and whisker plot, suggest that the higher PV frequency of the L2 data is as a consequence of a few PVs being used very often, rather than an overall tendency by L2 students to use PVs more often than do L1 students. In fact, the L2 frequencies in the table show the "rapidly diminishing word frequency" associated with Zipf's law (Brezina, 2018:44), starting with frequencies higher than that of the L1 data, and dropping to frequencies lower than that of the L1 data over the ten most used PVs (the highest drop in value being 49.79, and the lowest drop being 0.54). This "rapidly diminishing word frequency"

is not observed in the L1 data, where the drop in value only varies between 4.83 and 0.11. It should be noted that there is one noticeably large drop in frequency in the L1 data, between *grow up* and *end up* (47.64), but this does not adhere to the pattern for the rest of the data, suggesting that it is a random occurrence.

4.4.2.1 *Comparison of South African L1 / L2 and international most frequently used PVs*

An overview of South African student PV use compared to Chinese, British and American student PV use, as per Chen's (2013a) findings (see §2.9.7) provides further insight into South African PV use. As part of her research into PV use by Chinese students, Chen (2013a) did a comparison of the ten most frequently used PVs across various corpora. She used the Chinese learner corpus (consisting of academic essays), the LOCNESS-US, and the GS-UK. In order to have a standard against which PV use in academic registers could be measured, she also included information from the academic writing subcorpus of the Corpus of Contemporary American English (COCA-AC) and the academic writing subcorpus of the British National Corpus (BNC-AC).

The ten most frequent South African L1 and L2 PVs (given in Table 4.54) were compared to the five corpora used by Chen (see Table 4.55 below). The L1 PVs found in Chen's corpora are underlined, and the L2 PVs found in Chen's corpora are highlighted in light grey. South African L1 students show the greatest correspondence (50%) in PV use with the LOCNESS-US students. There is a 40% correspondence with the BNC-AC corpus, and a 30% correspondence with both the GS-UK and COCA-AC corpora. The Chinese student corpus shows the lowest similarity (10%). The PV use of native English speaking students in South Africa, therefore, most resemble that of American university students, albeit only by 50%. Interestingly, the correspondence with the two corpora that Chen (2013) included as standards against which performance could be measured, namely COCA-AC and BNC-AC (both of which comprise academic writing), is relatively low, at 30% and 40% respectively.

Table 4.55: Comparison of ten most frequently used PVs across six corpora, current research L1 and L2 breakdown

Chinese	LOCNESS-US	GS-UK	COCA-AC	BNC-AC	Present research L1	Present research L2
GO OUT	GO ON	SET UP	POINT OUT	CARRY OUT	CARRY OUT	GROW UP
SUM UP	TAKE AWAY	GO ON	CARRY OUT	POINT OUT	MAKE UP	END UP
GIVE UP	BRING UP	CARRY OUT	GO ON	GO ON	GROW UP	CARRY OUT
TURN OFF	GROW UP	FIND OUT	TAKE ON	SET UP	FIND OUT	MAKE UP
GROW UP	POINT OUT	RUN OUT	MAKE UP	SET OUT	END UP	FIND OUT
SET UP	END UP	BRING ABOUT	SET UP	MAKE UP	ACT OUT	POINT OUT
GO BY	GIVE UP	PUT FORWARD	TURN OUT	TAKE ON	BRING UP	PICK UP
CUT DOWN	BRING ABOUT	MAKE UP	BRING ABOUT	TAKE UP	SET OUT	SET UP
BRING ABOUT	FIND OUT	CUT DOWN	GIVE UP	BRING ABOUT	POINT OUT	FIGURE OUT
GO ON	SET UP	TAKE ON	PICK UP	TURN OUT	OPEN UP	GO OUT

Similar to the L1 students, South African L2 students also show a correspondence of 50% with the LOCNESS-AC corpus and 40% with the BNC-AC corpus. However, this is the only similarity between L1 and L2 use when compared to the other corpora. South African L2 use shows a correspondence of 50% with the COCA-AC corpus, and 40% with the GS-UK corpus. The lowest correspondence is with the Chinese student corpus, yet, at 30%, it is much higher than that of the South African L1 students. South African L2 students, therefore, also show a pattern of PV use that is similar to that of American students, and, although the similarity is only 50%, this similarity level is also repeated for the academic standard corpus, namely the COCA-AC (50%). It seems that the PVs used by South African L2 students show a higher correspondence with PVs used in American academic writing than does South African L1 PV use.

The overall picture that emerges is that South African student writing in general shows most similarity with American student PV use. It is possible that this is the result of the impact of American media on South African culture, which is not entirely unexpected, given the prevalence of PVs in informal registers. To what extent this is a result of the prevalence of American media (such as movies, music,

books and social media), would be an interesting area of study (as discussed in the concluding chapter, under 'Suggestions for further research').

4.4.3 Comparison of L1 and L2 ALT use

A preference for ALT or PV use was investigated for both L1 and L2 students, which will be reported on separately before a comparison is made between the two groups. In the course of the research into L1 ALT use, it became apparent that ALTs were preferred to PVs across all the L1 subcorpora. The process required the selection of suitable synonyms for the ten most frequently used PVs of each subcorpus, using Wordsmith Tools to extract all uses of the relevant ALTs, followed by the inspection of all extracted concordance lines for valid alternative uses for PVs. (See §3.8.2.5 for a discussion of the process of selecting ALTs.) In cases where the context of the concordance line indicated that the ALT was used in a different meaning sense to that of the PV, the concordance line was not included in the ALT frequency. Three ALTs, where possible, were investigated for each PV, because of the different meaning senses that might be associated with a PV. However, this was not always possible, as not all PVs have appropriate synonyms. In such cases, the synonyms delivered few suitable concordance lines, and low frequencies. Overall, the L1 students used 3 490.03 ALTs (relative frequency), compared to 519.57 PVs (relative frequency), which is a ratio of 6.7 ALTs for every PV (see Table 4.56). This indicates a preference for ALTs over PVs by L1 students.

Table 4.56: Comparison of L1 and L2 PV and ALT frequencies

Group	***Relative frequency of most used PVs****	***Relative frequency of ALTs****
L1	519.57	3 490.03
L2	632.84	2 575.91

*Per million words

As with the L1 subcorpora, the L2 extracted data showed that ALTs were preferred to PVs by the students represented in the L2 subcorpora. The same patterns that were observed for the L1 data were also apparent here, in that, when considered individually, ALTs were not always preferred to PVs, because of insufficient PV synonyms being available in certain cases. Nevertheless, overall, L2

students used 2 575.91 ALTs (relative frequency) compared to 632.84 PVs (relative frequency), which is a ratio of 4.07 ALTs to 1 PV. This is not as high as the ALT to PV ratio of the L1 students, yet still confirms a preference for ALTs over PVs by L2 students.

The ten most frequently used PVs to which the ALTs were preferred are also given in Table 4.56. (The frequencies presented in the table were collected from the individual results of the institutions and collated. They do not, therefore, represent the frequencies used in §4.3.3, as those frequencies represent a summary of the ten most used PVs overall.) The information presented here confirms the observation made in §4.3.1 that the L2 PV frequency is higher than the L1 PV frequency, by 21.8% ((*highest PV frequency - lowest PV frequency) / lowest PV frequency * 100*). Similarly, the L1 ALT frequency is shown to be 35.49% higher than the L2 ALT frequency. This supports the observation that L1 students are more inclined to use ALTs than L2 students, suggesting a more formal style of writing.

The L1 and L2 PV and ALT frequencies are illustrated in Figure 4.12, showing the higher PV use in the L2 subcorpus compared to the L1 subcorpus, and the higher ALT use in the L1 subcorpus compared to the L2 subcorpus.

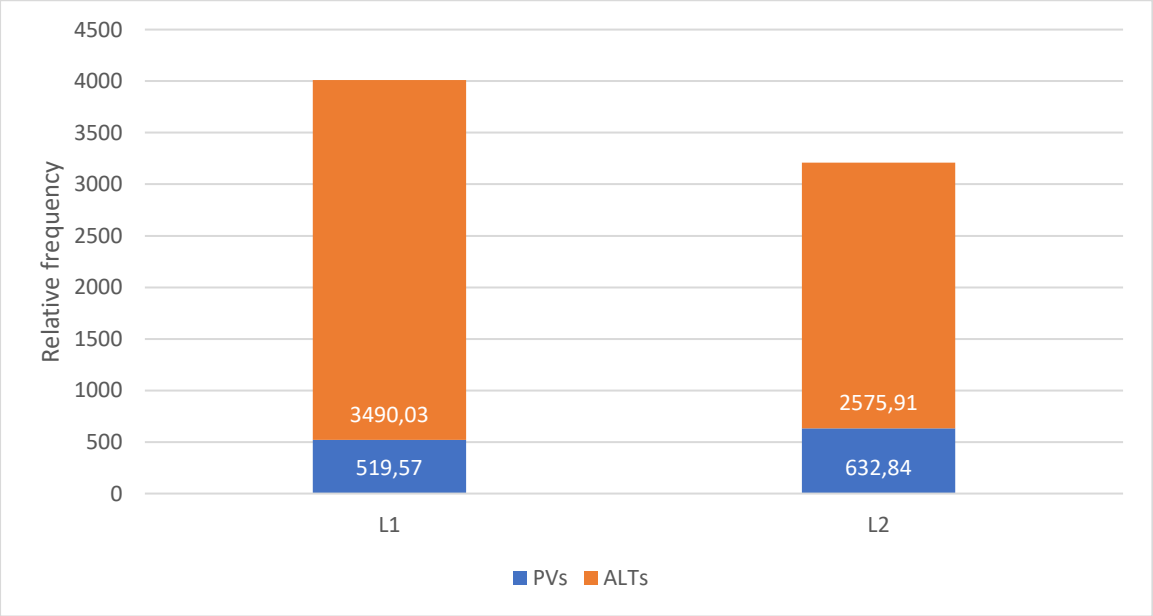

Figure 4.12: Comparison of L1 and L2 PV and ALT frequencies

L1 and L2 PV error frequencies will be examined next, as errors in PV use will have an impact on the apparent PV fluency suggested by one group's higher PV use.

4.4.4 Comparison of L1 and L2 PV error frequencies

The next aspect that will be considered is the error frequencies for the two groups overall. In the process of validating PVs by means of an inspection of various dictionaries and the scanning of concordance lines, any errors in the use of the PVs were noted and recorded in an Excel file. Both PV error tokens and types were reported on, as this indicated which specific errors had become entrenched in student writing. This was done per L1 and L2 subcorpus within each of the three institutions, and the data aggregated per L1 / L2 groups overall.

The types of PV errors that were recorded were not decided on independently, but were allowed to emerge from the data. These error types became apparent during the analysis of the first subcorpus (NWU L2), and all subsequent errors that emerged across the various subcorpora were found to adhere to the error types (of which there were five) that had originally been identified. The error types consisted of *redundant particle* (unnecessarily attaching a particle to a verb proper, thereby creating a PV when the verb on its own would have sufficed), *incorrect particle* (creating an incorrect *verb + particle* combination, as a different particle should have been attached to the verb proper to convey the correct meaning), *incorrect PV use* (using an existing PV in the wrong context), *non-existent PV* (using a PV for which no confirmation of legitimacy could be found in any of the formal and informal sources used), and *unconfirmed PV* (using a PV for which no confirmation of legitimacy could be found in any of the formal sources used, but which was referred to informally, suggesting a possible inclusion as a legitimate PV in the future). The PV error types *redundant particle* and *incorrect particle* are considered syntactic errors because of the incorrect use of a grammatical structure, whereas the PV error types *incorrect PV use*, *non-existent PV*, and *unconfirmed PV* can be categorised as semantic errors as they refer to meaning. No further PV error types surfaced during the research. On the other hand, the five error types did not feature in the error analysis of every subcorpus. For example, only

three error types occurred in the NWU L1 subcorpus. The distinction between *error type* and *error token* was useful as it indicated the recurrence of particular error type examples, which suggested that these incorrect uses of PVs might, to an extent, have become incorporated into student writing.

As shown in Table 4.57, the L2 relative error type frequency is higher than the L1 relative error type frequency by 61.86% (50.08 – 30.94) / 30.94 * 100), Thus, it can be surmised that L2 students were more inclined to make mistakes in the use of PVs than the L1 students. The L2 relative error token frequency is also higher than the L1 relative error token frequency by 67.9% (73.26 – 43.63) / 43.63 * 100), which further suggests that there was a greater likelihood of the same PV error occurring more than once (for example, the invalid PV *issue out* appearing 12 times in the NWU L2 subcorpus). A higher error token frequency than error type frequency indicates the possibility that certain errors might have become entrenched. However, the similarity of the ratios of error type frequency to error token frequency for the two groups (with 1:1.41 for the L1 group, and 1:1.46 for the L2 group) indicates that the entrenchment of errors occurred to the same extent for both groups.

Table 4.57: Comparison of L1 and L2 error frequency

Group	***Relative error type frequency****	***Relative error token frequency****	***Ratio of error types to error tokens***
L1	30.94	43.63	1:1.41
L2	50.08	73.26	1:1.46

*Per million words

Figure 4.13 illustrates the difference in error type and token frequencies of the L1 and L2 subcorpora.

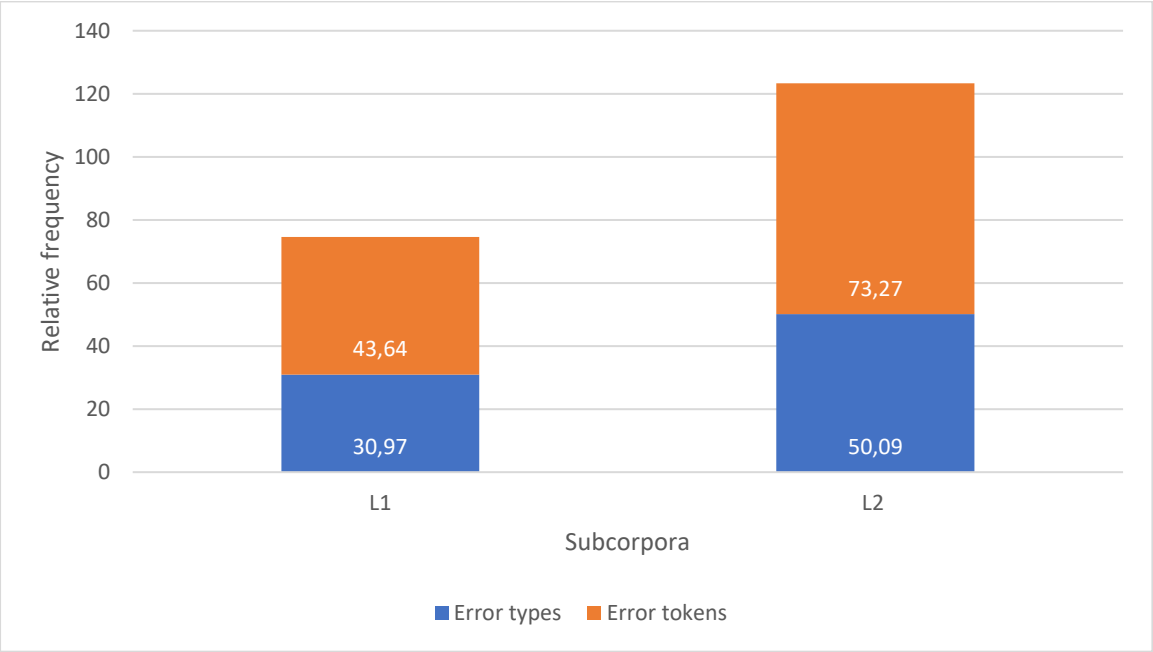

Figure 4.13: Comparison of L1 and L2 error frequencies

It was also of interest to see whether the PV errors that students made fell into syntactic or semantic categories, and whether L1 and L2 student performance differed in these areas. Table 4.58 shows the difference between L1 and L2 students regarding these error categories.

Table 4.58: L1 and L2 comparison of adherence to syntactic and semantic norms

Type of error	***Type of PV error***	***Relative error type frequency****		***Relative error token frequency****	
		L1	***L2***	***L1***	***L2***
Syntactic	*Redundant particle*	10.81	18.57	15.23	28.25
	Incorrect particle	3.63	6.53	6.68	8.42
	Subtotal	**14.44**	**25.10**	**21.91**	**36.67**
Semantic	*Incorrect PV use*	7.27	15.35	8.84	25.07
	Non-existent PV	6.09	6.35	7.97	7.22
	Unconfirmed PV	3.14	3.28	4.91	4.30
	Subtotal	**16.5**	**24.98**	**21.72**	**36.59**
	Total	**30.94**	**50.08**	**43.63**	**73.26**

*Per million words

Relating these two categories of error to student writing shows that L2 students had the highest relative token frequency for syntactic errors (36.67), as well as the highest relative token frequency of semantic errors (36.59). This is not unexpected, given the higher overall relative error token frequency for the L2 students given in Table 4.57.

The L1 error type frequency for syntactic errors makes up 65.91% (*error type frequency / error token frequency * 100*) of the L1 error token frequency, which suggests that exactly the same error was repeated in more than half the cases. For example, *issue out* (see §4.2.1.1.4) contains a redundant particle, and adds one count to the redundant particle error type category. On the other hand, if this particular combination (*issue out*) of the redundant particle error type category should appear 12 times, then a count of 12 will be added to the redundant particle error token category, which will result in the error token frequency being higher than the error type frequency for this particular error category.

Likewise, the L2 error type frequency is 68.45% of the L2 error token frequency. On the other hand, the L1 error token frequency for semantic errors comprises 75.97% of the L1 error type frequency, which is 10.06 percentage points higher than its syntactic percentage of types to tokens. A high percentage of semantic error examples, therefore, seem to have been repeated. This is somewhat unexpected, given that these are first-language speakers, and, as such, would have encountered the use of the PV in context, and learnt the appropriate meaning. In contrast, the L2 error token frequency for semantic errors includes 68.27% of the L2 error type frequency, which is roughly the same as its syntactic percentage of types to tokens.

It is interesting to note that, as far as syntactic errors are concerned, *redundant particle* was the error with the highest frequency for both the L1 and L2 groups. Similarly, for semantic errors, *incorrect PV* use was the most frequent error for both groups. This suggests similarity in PV error patterns among the L1 and L2 students, although the error frequencies differ.

In the next section, the various aspects regarding L1 and L2 PV use within and across institutions, as well as overall PV use by L1 and L2 students, as discussed previously in this chapter, will be summarised.

4.5 Findings

This research study investigated PV use by South African students with the primary goal of contributing to research into L2 English proficiency in South African education. However, patterns of L2 language use should not be viewed in isolation in the South African context, as the language landscape is somewhat unique. Thus, L2 student PV use was examined in conjunction with L1 student PV use to obtain a comprehensive view of the use of this feature within the South African environment.

PV use was examined from various angles, as frequency alone was likely to result in a skewed view of actual PV competence. Therefore, besides measuring overall PV token frequency, PV type frequency was also considered, as this gave an indication of the number of PVs that were used repeatedly, compared to PVs that were used on a once-off basis. Information on PV token and type frequencies within a subcorpus allowed for an assessment of PV distribution, which could then also be compared between L1 and L2 groups. The ten most frequently used PVs were extracted for particular comparison, to assess whether frequently used PVs differed among the L1 and L2 groups, and across institutions. As PV avoidance was determined by means of a noticeable preference for ALTs in global PV research, relevant ALTs were selected for the ten most frequently used PVs and their frequencies extracted from the subcorpora. The resultant ALT frequencies were then compared to their PV counterparts to assess whether student preference was for ALTs or PVs. Finally, non-standard PVs that had surfaced during the PV validation process were reported on, as such instances of PV use will have an impact on an assessment of PV proficiency.

L1 and L2 PV use was first examined separately within each of the institutions (NWU, UP and WITS), and then compared. Once this information was available, PV use could be compared across the three institutions using the same parameters as those that had been applied within each institution. While the ultimate aim was to determine overall L1 and L2 patterns of PV use, examining PV use within and across institutions allowed for nuanced observations of PV use. The final step was to compare L1 and L2 PV use, as this formed the main focus of the research questions.

While it had been surmised that the PV use observed at a university would coincide with its international ranking (*Times Higher Education Sub-Saharan Africa University Rankings* 2024), this assumption proved to be invalid. WITS students showed the greatest inclination to use PVs, and the greatest familiarity with certain PVs of the three institutions. The UP students showed the least inclination to use PVs, and the greatest inclination to use particular PVs once only. UP students were also more inclined to use ALTs. Lastly, in a comparison of incorrect PV use as a proportion of PV token frequency, UP students were shown to be most likely to use PVs incorrectly. Separately comparing the L1 and L2 PV use by the institutions supported the findings discussed above. Therefore, considering these results, WITS could indeed be said to align with its position as the highest ranking of the three institutions represented in this research. However, UP appeared to display the weakest PV proficiency although it is ranked second of the three institutions. Therefore, the level of PV use at a university does not necessarily coincide with its ranking.

Differences in overall PV use between the L1 and L2 subcorpora may be summarised as follows. The L2 students represented in the corpora displayed a higher inclination to use PVs than did L1 students. On the other hand, the L1 students were more inclined than the L2 students to use PVs more than once. Examining PV dispersion indicated that, though a limited number of PVs were used with relative consistency by both the L1 and L2 student groups, the majority of PVs in the two groups was used once only. This is supported by an investigation of the most used PVs of the L1 and L2 groups across the three institutions, which indicated a steady drop in frequency for the top ten PVs.

The top ten most frequently used PVs of the L1 and L2 groups warrant further comment. Seven of the ten most used PVs appeared in both the L1 and L2 lists, although the frequencies at which they occur differ. The fact that the same PVs should be used by both groups of students suggests that particular PVs are well known to South African students generally, and, by extension, that these PVs are part of the South African landscape. Basing the teaching of PVs on these frequently used PVs could serve as a convenient entry point as learners would already be familiar with some of them, though unaware of their use and structure. Important aspects of the PV can then be introduced, such as its syntactic structure, emphasising its separability, and its meaning, which should include a discussion of the variety of meaning senses of some PVs, as well as how meaning is affected by category of PV, ranging from literal to idiomatic. PV use across registers should be covered, which also necessitates a discussion of ALTs, with examples of contextual sentences as would be found in concordance lines. PVs. (Some of these aspects are elaborated on below.)

An investigation into ALT or PV preference by L1 and L2 students demonstrated that ALTs were preferred to PVs by both groups. However, the pattern of ALT preference differed, in that the L1 students showed a substantially greater inclination (by 35.49%) to use ALTs than the L2 students. This notable preference for ALTs over PVs by the L1 students intimates a more formal vocabulary than that of the L2 students, given the formality normally associated with one-word “Latinate” verbs, in contrast to PVs (see §2.5). An informal writing style should not necessarily be viewed as deficient, as Chen’s (2013a) comparison of British and American student writing indicated. Nevertheless, should a more formal academic writing style be preferred in South African educational circles, L2 learners might be well served if they were to be provided with a list of PVs and their ALTs, such as contained in the tables that resulted from the examination of L1 and L2 ALT use within the various institutions. This would provide an opportunity for indicating which PVs should preferably be replaced by an appropriate ALT in academic writing, while also demonstrating that not all PVs have suitable synonyms. Furthermore, students will be made aware that some PVs have a variety of meaning senses, influencing the choice of an appropriate ALT. This is an important aspect of PV use that learners will be required to master.

Another aspect of PV use that was considered in the research was PV error frequency, as PV frequency cannot be viewed without a consideration of incorrect PV use. An evident inclination to err in the use of PVs, be it syntactically or semantically, would cast doubt on the PV proficiency suggested by high PV use. Indeed, while the L2 PV frequency proved to be higher than that of the L1 PV frequency, the L2 PV error frequency was likewise higher than the L1 PV error frequency. This suggests that L2 students are less familiar with the rules and norms governing the use of PVs than L1 students. The identification of the types of PV errors that occurred within the various subcorpora provide teachers with a useful starting point for guiding learners in correct PV use, as the error information is not based on intuition, but on authentic data drawn from the concordance lines of actual student texts. Using the actual examples of PV errors provided in the various subsections would add authenticity to their teaching practice.

Ultimately, though, overall South African student PV use in the corpora investigated has been shown to be low when compared to PV use in international corpora (see §4.3.2). Therefore, it can be asserted that PV use does not feature strongly in South African student writing. As this is true for L2 students, and, more specifically, for L1 students (which indicates that PV avoidance is not a factor), it suggests that South African students favour a formal style of academic writing, possibly as a result of an emphasis on register at school. Incorrect PV use was found to be higher for L2 students than for L1 students, which might point to the use of a grammatical structure without adequate knowledge of the use of the structure. On the other hand, the overall presence of PV errors was notably low (see §4.3.5) which suggests that South African students, whether L1 and L2, have a good grasp of PV use, albeit possibly intuitively. Nevertheless, learners will benefit from instruction regarding the proper use of PVs in academic writing.

4.6 Conclusion

This chapter covered the analysis of data pertaining to PV use in the SAMuLCAT and WITS corpora. These corpora contain the student writing of three South African higher education institutions: NWU, UP, and WITS. The data of each institution were subdivided into L1 and L2 texts because the focus of the research was to report on PV use in these subdivisions. L1 and L2 PV use was compared within each institution, and then across the institutions. Subsequently, an overall comparison was made between the L1 and L2 groups, which was the main focus of the research. By means of these comparisons, patterns of PV use emerged and could be commented on, thereby establishing the groundwork for a discussion of the findings.

The chapter that follows considers the contribution of this research, and makes pedagogical recommendations based on the findings discussed here. The limitations that might have had an influence on the outcome are also outlined. Finally, suggestions are made with regard to possible future research into PV use in the South African context.

Chapter 5: Conclusion

5.1 Introduction

In the concluding chapter, the aim of the research will be reviewed, and the findings discussed. Thereafter, the contribution that this study hopes to make will be considered. Limitations that might have had an impact on the research will also be acknowledged. The chapter will conclude with suggestions for future research.

5.2 Review

Global research into PV use suggests that L2 students struggle to attain competence in the use of this grammatical feature which is so prevalent in the English language. This research study sets out to discover whether this trend was also observable among South African L2 students, and, in addition, to examine the general patterns of PV use by L2 students. In order to obtain as clear a picture of L2 student PV use as possible, it was also necessary to widen the investigation to include L1 student PV use. Assessing L2 PV use in isolation would have provided limited information, from which it would not have been possible to make valid inferences.

Two corpora, SAMuLCAT and WITS, were used in the research. Both corpora consist of student writing, which made it possible to use authentic data in the research. Whereas the WITS corpus contained data from only one academic institution, two academic institutions, NWU and UP, were available in the SAMuLCAT corpus. The corpora were separated into L1 and L2 subcorpora because of the nature of the research aims. Student academic writing at these three institutions formed the basis of the research.

Investigating PV use by L1 and L2 students was approached from various angles in order to gain the most complete picture possible on this topic. The most crucial aspect to consider was PV frequency, as this formed the core of the research, from which all other aspects of PV use flowed. Not only was the total number of PV tokens extracted from each subcorpus, but also the total number of PV types

(or base forms), in order to establish whether there were PVs that were used with a degree of consistency, or whether, alternatively, PVs were only used incidentally. In addition, the distribution of PVs in individual subcorpora was examined. Thereafter, the ten most used PVs per subcorpus were extracted in order to reveal the specific PVs that were used with regularity by the students represented in that subcorpus. This also allowed for the particular PVs used by students across universities to emerge, once the subcorpora were compared. Student preference for either PVs or ALTs was then established, as a noticeable preference for the latter would be expected to have an influence on PV use. Finally, the PV errors that were observed in student writing were reported on, to establish whether students adhered to syntactic and semantic norms in their use of PVs.

Previous research suggested that L2 students used PVs less than L1 students did, but that this observation was sometimes, but not always, negated by the influence of a mother tongue with a similar construction, or by a high level of English proficiency. Other studies also indicated that L2 students were likely to use more ALTs than did L1 students, and make more mistakes in the use of PVs. The findings of the present research, and the degree to which the findings adhere to these expectations, will be discussed in §5.2.2.

5.2.1 Aims and research questions

The main focus of the study was to establish the patterns of PV use by South African L2 students. Thus, data on L2 PV use needed to be extracted from the relevant corpora and subcorpora. However, in order to make valid observations, various other perspectives had to be taken into consideration. For this reason, further subordinate research aims were formulated. These research aims were intended to determine how L1 and L2 students used PVs in their writing, followed by an investigation into whether there was a difference in patterns of PV use at different universities. The extracted data could then be used to assess the overall difference in PV use between L1 and L2 students. It was hoped that identifying these patterns of use would assist in teaching correct PV use to L2 students.

Research questions were devised to investigate these research aims, and to direct the research. The first set of research questions was used to garner information about the pattern of PV use in L1 student writing:

1a. Which phrasal verbs are predominantly used by L1 students?

1b. To what extent do L1 students show a preference for phrasal verbs or for one-word alternatives?

1c. To what extent do L1 students adhere to syntactic and semantic norms in their use of phrasal verbs?

The next set of research questions was then used to garner information about the pattern of PV use in L2 student writing:

2a. Which phrasal verbs are predominantly used by L2 students?

2b. To what extent do L2 students show a preference for phrasal verbs or for one-word alternatives?

2c. To what extent do L2 students adhere to syntactic and semantic norms in their use of phrasal verbs?

The next research question was created to assess the differences in PV use between L1 and L2 students.

3. What are the main differences in the use of phrasal verbs by L1 and L2 students?

A research question was also created to investigate whether PV use across universities differs according to university rankings.

4. Does the occurrence of PV use differ across universities according to the ranking of each university?

The final two questions focus on the contribution that this study hopes to make to the teaching of PVs.

5. How could raising awareness of phrasal verbs and their alternative one-word verbs help students make appropriate choices in their academic writing?

6. How could reviewing errors in the use of phrasal verbs serve to guide students on how to use phrasal verbs correctly?

5.2.2 Main findings

The problem statement in §3.2, which is based on the findings of global research on the topic, suggests that research conducted in the South African context is likely to show that South African L2 students experience problems with the use of PVs. This is in keeping with the findings of the various research studies discussed in the literature review (Chapter 2) that suggest that L2 attainment of competence in English is negatively influenced by unfamiliarity with and avoidance of the use of PVs (Dagut & Laufer, 1985:4; Hulstijn & Marchena, 1989; Laufer & Eliasson, 1993; Liao & Fukuya, 2004). Consequently, one of the assumptions underlying the study was that similar patterns of PV avoidance would be observed in South African L2 student writing.

Nevertheless, in contrast to the findings mentioned above, the L2 students investigated in this research were found to use PVs more often than the L1 students. Although the difference was not noteworthy, the result is unexpected as it differs to a large degree from the results reported in other research. This is not without precedent, however. As noted in §2.7.7, German students (Riguel, 2014:114; Waibel, 2007; Zhou & Wang, 2024:4), as well as Malaysian L2 students (Zhou & Wang, 2024:4), were found to use more PVs than do native speakers. Similarly, research conducted by Chen (2013b:433) indicated no difference in PV use between British and Chinese students, and Waibel (2007) found no difference in the frequency of PV use when Dutch and Polish students were compared to a native corpus.

The fact that South African L2 students were found to use PVs more often in their writing than did L1 students, therefore, is not an isolated case. It is worth noting, however, that the South African situation is unique, in that, in this case, the L2 students are attempting to acquire the *lingua franca* of their native country. It can be surmised that most L2 students in South Africa are likely to have been exposed to English from an early age, and will have grown up surrounded by some form of colloquial English, considering the presence of radio, television, cinema, newspapers, and social media in everyday life. This is reflective of the "frequent English exposure out of class" mentioned by Zhou and Wang (2024:4). Thus, the profile of the L2 student in South Africa is different from that of the L2 students discussed in §2.9. In contrast to students learning English as a foreign language, where each new grammatical concept must seem alien, L2 learners in South Africa, being surrounded by the language they are learning, will have picked up some of these concepts subconsciously. Whatever the reason for the higher use of PVs by L2 students compared to L1 students shown in this research, it is clear that the use of PVs by L2 students within the South African context does not reflect the deficiency reported in international research.

Liao and Fukuya (2004:214) observed that L2 students in an immersive environment are more likely to be familiar with PV use due to its informal nature. Schmitt and Redwood (2011) also found that PV frequency (that is, extensive exposure to PVs in an L1 environment, via reading and popular media) had a positive influence on PV acquisition. Likewise, Waibel (2007:167) found that advanced German students were able to master PVs primarily by informal means, such as "increased input of authentic English through the media". Waibel (2007:167) also included "a prolonged stay in an English speaking country" as an example of an informal means of learning to use PVs. Similarly, Siyanova and Schmitt (2007) reported that an extended stay in a country where English is the native language resulted in increased PV use. Therefore, it is to be expected that most South African L2 students will have some familiarity with frequently encountered PVs, having been exposed to English in their everyday lives, but without formal training in identifying the PV as a unique grammatical structure. That there should be an overlap in the most frequently used PVs by L1 and L2 students (as observed in §4.4.2) seems to

support the idea of the ubiquity of English, as this overlap shows that all students, whether L1 or L2, are possibly confronted by certain PVs on a daily basis. Formulaic sequences (one of the L2 learning techniques discussed in Chapter 1) could possibly be relevant to the acquisition of PVs (Hummel, 2014:142), as PVs might very well be learned as formulaic sequences by L2 students in the South African context, given the prevalence of English in the linguistic landscape of the country. As formulaic sequences are thought to be learnt as units, the multipart form of the PV might not be questioned by the L2 student. This, of course, is very likely to result in a lack of knowledge about the rules governing these structures, thereby leading to, at times, inappropriate or incorrect use. It should be noted that such inappropriate or incorrect use was not confined to L2 student writing, and that, in fact, similar PV error patterns were observed among the L1 and L2 students (see §4.4.4). Furthermore, PV errors did not make up a notable percentage of any of the subcorpora, whether L1 or L2 (see Table 4.50).

In the South African context, it might justifiably be speculated that the formal teaching of English at school level will have less influence on PV use than the informal exposure described above. On the one hand, no evidence could be found of this construction being taught formally in South African schools (refer to the *English Handbook and Study Guide* (Lutrin & Pincus, 2007), a reference book for South African learners, and the *Learning and teaching support materials* of the Department of Basic Education (2021)). Morales (2000a:6) is of the opinion that "[m]any teachers believe it is not worth investing time in teaching informal expressions, such as phrasal verbs". It is certainly true that classroom time constraints do not allow for instruction on all aspects of language (Morales, 2000a:6). Yet L2 students need to have some knowledge of even informal language structures that they might encounter in informal settings. Moreover, it is important to remember that L1 students also used PVs incorrectly at times, which might suggest that, though their relatively few errors are the result of intuitive knowledge, they are not entirely conversant with the structure. On the other hand, considering the complicated and sometimes arbitrary tests needed to validate PVs, PV instruction could "seriously impair the learning of phrasal verbs", and lead to "students' avoidance of phrasal verbs" (Darwin & Gray, 1999:67, 75). Chen (2013b:98) speculates that "[t]oo much explicit input in

classroom learning might lead to the opposite result", so that, once students become aware of the difficulty inherent in PV use, they might be inclined to avoid the construction. Thus, the possibility exists that L2 students will be less likely to use PVs to the same extent that they do at present were they to receive formal instruction on the use of PVs. Consequently, it is important to make the teaching of PVs as accessible as possible in order to facilitate, rather than to obstruct, knowledge of the correct use of PVs. (Suggestions for possible teaching practices are discussed in §5.4.)

A further point to consider refers to the difference that was observed by Chen (2013a:435) between the PV use of British and American students (see §2.9.7), with British students employing a more formal style of writing, and American students a more colloquial style. As both these groups are native speakers, the problems with PV use encountered by L2 speakers do not apply here. Chen (2013b:95) suggests that a difference in writing style may account for the variation in PV use reported for the two groups, in that "American written genres show more informal colloquial and interactive features". As mentioned previously, British student writing, in contrast, is inclined to be more formal. Zhou and Wang (2024:4) likewise report that British students use fewer PVs than American students, and that they also have a "clearer genre awareness by using PVs less in academic papers". Applied to the South African context, the difference in PV use might be ascribed to a more formal writing style employed by L1 students, compared to L2 students. The low PV error frequency noted for the L1 students supports this theory, as it suggests competent PV use. This is also supported by the higher L1 PV type frequency, which shows the use of a greater variety of PVs by the L1 students compared to the L2 students. These aspects challenge the idea of competent L2 PV use that seemed to be suggested by higher L2 PV use overall. Rather, this might be the result of a limited, or less sophisticated vocabulary (Cooper, 2017:149) ("sophisticated", in this case, referring to mother tongue fluency and so an understanding of the more complex features of vocabulary items such as connotations and appropriate use of synonyms according to context), in contrast to that of the L1 students, as suggested by the higher L2 PV error frequency, and lower PV type frequency. Given the points raised here, it might also reflect a lack of awareness by L2 students of the appropriate register to use in academic

writing. On the other hand, Van Rooy and Terblanche (2006:178), in their research on the Tswana Learner English Corpus (TLEC), found Tswana student writing to be "more colloquial in style", which suggests that other aspects might also influence writing style.

To further understand South African PV use, the use of ALTs in preference to PVs should also be considered. Global research has indicated a preference for ALTs over PVs by L2 students, as they attempt to avoid a grammatical structure with which they are unfamiliar (Laufer & Eliasson, 1993; Siyanova & Schmitt, 2007). While an L2 preference for ALTs was likewise observed here, the higher L2 PV use, compared to that of L1 students, suggests that avoidance was not a factor in this case. A similar L1 preference for ALTs over PVs, likewise, is in contrast to the PV use pattern observed elsewhere (Hulstijn & Marchena, 1989:242; Laufer & Eliasson, 1993:36). The L1 students cannot be said to be avoiding the PV structure because of unfamiliarity, which further negates the idea of ALT preference in this case suggesting PV avoidance. In the South African context, the L1 preference for ALTs is also noticeably higher than that of the L2 students. This further supports the idea of a more formal style of writing being used by the L1 students, and suggests that they "have a better grasp of academic vocabulary", compared to that of the L2 students (Cooper, 2017:149).

The overall low PV use reported at the institutions investigated in this research (see §4.3.2) necessitates a final comment. This indicates that, in general, South African students do not use PVs widely in their academic writing. This might suggest that these students, when engaged in academic production, are, in fact, inclined to adopt a formal style of writing. However, this can only be confirmed if investigation of a South African student speech corpus were to indicate high PV use, proving that students use PVs more extensively in a colloquial and informal environment.

The pedagogical implications of the information discussed here will be given in §5.4, after a brief discussion on the contribution that this research hopes to make in the South African educational environment.

5.3 Contribution

PV use by L2 students has been well researched globally. Yet, although the PV is seen as a grammatical structure that is generally problematic for non-native speakers, and is furthermore an indicator of fluency in English, no major studies are on record as having been conducted into its use by L2 students in South Africa. This study does not claim to fill this gap entirely, but hopes rather to make an initial contribution from which more research and insights will flow.

As discussed previously, the results suggest that L2 PV use differs somewhat from that observed elsewhere. This fact has pedagogical implications, in that current material used by schools does not appear to address the issue of PV use. Whether PV use by learners needs to be encouraged (resulting in the American pattern of use, as reported by Chen (2013a)), or discouraged (resulting in the British pattern of use (Chen, 2013a)) (see also §5.4), is a decision that does not fall within the domain of this research, but rather within that of the relevant educational authorities. Nevertheless, it should be pointed out that teaching learners to avoid the use of the PV altogether will result in their English sounding unnatural, as the PV is "an important component of English vocabulary" (Garnier & Schmitt, 2016:30). Therefore, regardless of the stance taken towards PV use in academic writing, student knowledge of this structure is essential. For this reason, evidence of differences in PV use between South African L1 and L2 students should prove useful to decisionmakers.

5.4 Pedagogical implications and recommendations

As discussed in the literature review (Chapter 2), there are varying opinions as to the suitability of PV use in academic writing. The conclusion was drawn that the discretionary use of PVs was acceptable, but that informal PVs, especially PVs that might be considered to be slang, were to be avoided, as the inclusion of such PVs compromises the quality of academic writing.

The results discussed in §5.2.2 suggest that South African L2 students are less likely to avoid PVs in their academic writing than are L1 students. Furthermore, the investigation into the ten most frequently used PVs by L2 students showed that these PVs are similar to those used by the L1 students.

Consequently, it can be surmised that L2 students use acceptable PVs in their writing. However, the ease with which L2 students use PVs (as indicated by the higher PV use by these students) might suggest a more casual attitude to PV use in academic writing than that employed by L1 users. This is confirmed by the higher use of ALTs by the L1 students, suggesting a more formal writing style. Knowledge of PV use is necessary for an understanding of the appropriateness of its inclusion in academic writing. Therefore, though the intention is not to suggest that L2 students need to be "schooled in dominant forms of language and literacy" (Boughey, 1998:171), instruction at an educational level in the appropriate use of PVs in academic writing might benefit all students.

In discussing the findings of the data analysis (§4.5), some suggestions were made for applying the results of this research to the teaching of PV use. While this study has shown that South African students do use PVs, they possibly (as there is no clear sign of the inclusion of the PV in the curriculum) do not know that the structure they are using is called a phrasal verb, that it functions in a certain way, and that some forms of this construction would be considered slang.

An initial introduction by means of already familiar PVs would ease learners into learning about the concept of the PV, and what a PV entails. In this regard, the table of the most frequently used PVs provides a convenient list for use in the classroom, as it presents PVs that were used across the three institutions, and, thus, are likely to be familiar, at least to some degree, to learners, while also representing "authentic language use" (Garnier & Schmitt, 2015:649). Furthermore, with the use of a list such as this, learners can be made aware of the two-part structure of the PV, which should lead to a discussion of its separability. The tables that provide information about the ALTs for the most frequently used PVs could similarly prove useful for teaching various aspects regarding the PV. For example, this information is useful for introducing PV meaning, as the choice of ALTs will indicate that certain PVs have no suitable synonyms, and that certain PVs have a range of possible meanings, or meaning senses. The subject of PV meaning will also facilitate a discussion on the differences inherent in literal and idiomatic PVs. The use of PVs across registers might also be approached here, with a

clarification of inappropriate PV use, such as slang forms. Correct PV use might be usefully illustrated and reinforced by a reference to the PV errors examined in the study, especially as authentic examples have been made available for the use of the teacher.

Corpus linguistics has made it possible for researchers to have a better understanding of student PV use (Chen, 2013b:91; Gardner & Davies, 2007:342). Yet, despite, our increased knowledge about the PV, there is still no clarity as to the most effective teaching methods for its acquisition. However, it is evident that whatever PV instruction is provided, it should help rather than hinder (referring to the hindrance to learning that might result from the complicated nature of PV identification), as was mentioned in §5.2.2. Badem and Simsek (2021:65) propose that "an explicit way of teaching should be considered by language teachers", as a "conscious awareness" is needed to grasp the concept of the PV's construction. Using real-life situations is one such method suggested by Badem and Simsek (2021:65). Zhou and Wang (2024:19) propose that "the instruction of PVs should be accompanied by ample authentic L2 input". Educators are also urged to consider language differences between the L1 and L2, especially when the L2 does not contain a similar construction. However, this could be problematic in the South African environment, as many languages might be represented in a classroom, and knowledge of the different languages might not be readily available to the teacher. Haugh and Takeuchi (2023:657) further point out that teachers themselves might also be insufficiently versed in the intricacies of PV use. If instruction in PV use has not been part of the South African school curriculum up to this point, it would be entirely reasonable to find that the teacher needs support in acquiring the necessary skills for teaching PV use.

Alternative educational methods have been found to be useful for teaching complicated grammatical structures such as PVs. Such methods are in keeping with the informal methods of acquiring PVs mentioned earlier. For example, Saraswati (2024:6) suggests that the use of digital technologies (such as watching English movies, generally a pleasurable pastime) can help young people "to be more independent in their learning", and, furthermore, can "facilitate the development of a global outlook".

In his research, Alhatmi (2023:9) found that providing contextual cues helped L2 students in selecting the correct PVs to use, and that "movies and TV shows are characteristic of being full-fledged contextualized language sources". While the principle of using everyday activities as teaching opportunities is sound, there is no evidence that these activities do, in fact, result in improved PV use (Saraswati, 2024:83). On the other hand, Roohani and Heidari Vincheh (2023:376) report that using gaming applications on mobile devices in the teaching of language has shown success, although it is still a reasonably new tool in this field. Making use of an environment that is familiar to many learners, such as a game on a mobile device, can be an enticing way of teaching language, motivating learners, and encouraging self-regulated learning. In fact, Roohani and Heidari Vincheh (2023:395) found that a game-based method of instruction proved to be the most successful in PV acquisition, compared to social media and classroom-based teaching methods. However, it should always be kept in mind that, if L2 students are simply regurgitating the PVs that they hear through the entertainment media, without being taught how to discern between academically acceptable and colloquial PVs, they cannot be expected to improve their PV use when writing academically (Zhou & Wang, 2024:19). Occurrences of unconfirmed PVs, as discussed in Chapter 4, might very well be the result of informal language use learnt via social media.

On a more formal level, Zarifi and Mukundan (2014:52) have found that many textbooks aimed at teaching English to foreign- or second language learners are inadequate for teaching PVs, resulting in many of the problems that L2 learners encounter. Consequently, it is their opinion that learning materials need to be updated in order to make PVs more accessible, for example, as Waibel (2007:166) suggests, by incorporating the PVs used by L1 speakers. (See the discussion above on including frequently used PVs in teaching.) Gilquin (2015:82) maintains that such improvement of pedagogical materials could be expected to "enhance learners' knowledge of phrasal verbs", and it is highly likely that this applies to the South African context as well. Nevertheless, Waibel (2007:166) maintains that textbooks sometimes "clearly reflect the view of material designers and teachers that the teaching of

phrasal verbs is not worthwhile". This might be as a result of the fact that, because of their colloquial nature, they are so often deemed to be inappropriate in academic writing.

Besides the production of adequate textbooks, it is also necessary to provide students with information about the most frequently used PVs. According to Liu and Myers (2020:405), there are over 10 000 PVs in English. However, as confirmed by the current research, only "a small number of the most frequent PVs account for the majority of the uses of PVs in English" (Gardner & Davies, 2007). As discussed above, providing a list of the high-frequency PVs generated by an investigation of South African L1 and L2 PV use will provide a useful introduction into PV use for learners, because of the possible familiarity of some of these PVs. In addition, it might also prove useful to construct a list of high-frequency PVs commonly found in academic discourse to share with students.

5.5 Limitations of the study

Limitations that could have affected the results of the study should be acknowledged. Thus, possible limitations will be described here. First of all, the marks of the students represented in the SAMuLCAT corpus were not available to the researcher, which meant that performance could not be linked to PV use. Consequently, it was not possible to establish whether there was a link between frequency of PV use in academic writing and strong academic performance. This would be an interesting and worthwhile topic for future research.

Student educational background was also not considered in this study, as such information was limited to school location in the case of NWU and UP, or not available, in the case of WITS. Consequently, no differentiation was made between students based on quality of schooling. Other information that could have provided useful categorisation of the results is, for example, exposure to English and proficiency level.

A further issue that could have influenced results concerns the accuracy of the metadata that were supplied by the students. In the present study, information regarding a student's home language was

of particular importance because of its role in the separation of the corpora into L1 and L2 groups. In the case of the SAMuLCAT corpus, this would have been the *school home language* field, and, in the case of the WITS corpus, the *1st/2nd language* field. These fields indicate the language with which the student identified while at school. According to Cooper (2016:98), there are two possible reasons for the misrepresentation of one's home language. Firstly, she suggests that English might be viewed as "a marker of prestige within the context of English-medium universities" (Cooper, 2016:98). Secondly, students from multilingual households who consequently attend school in English might, as a result, decide to identify as English-speaking. Such misrepresentation of one's home language is not unique to South Africa, and has been reported by other researchers (Saville-Troike & Barto, 2017:11). The home language indicated by a student is taken at face value, and cannot be verified. It is hoped that any discrepancies in this regard will not have greatly influenced the results.

Furthermore, because the qualitative analysis of the research is necessarily manual in nature, it is to be expected that errors might have occurred in the process, regardless of the care taken by the researcher to prevent such errors. It is hoped that the possibility of error was adequately acknowledged and discussed in Chapter 4. The following serves as an example. One of the qualitative analysis aspects used in this study was the selection and extraction of ALTs. In research studies using classroom tests to investigate PV use (see §2.9), the choice of ALTs could be controlled through preselection by the relevant researchers, whereas this was not possible in the present study as it makes use of existing corpora. While a logical process was used to investigate ALT use, it is by no means claimed to be foolproof. Nevertheless, the researcher is of the opinion that results emanating from real-world data (in this case, corpora consisting of actual student writing) provide valuable and necessary information regarding PV use.

In the next section, some suggestions are offered for future research into the topic of PV use in the South African environment.

5.6 Suggestions for further research

In their research into *Native and nonnative use of multi-word vs. one-word verbs*, Siyanova and Schmitt (2007) found that, apart from avoidance, the PV input that L2 speakers receive through exposure to the L1 in various forms might also have an effect on PV use. Chen (2013a:429) maintains that "acquisition and production of phrasal verbs are dependent on their frequencies of occurrence in learning materials". It would thus be useful to investigate whether this dependence manifests in a South African context by incorporating into a future study the academic texts, articles and learning materials to which the participants in this study were exposed. Such a study might provide useful information regarding the use of PVs in the South African academic discourse, and, at the same time, indicate which of the various forms of published writing proved to have the greatest influence on students' inclusion of PVs in their writing. Furthermore, the genre of the academic texts from which PVs were drawn for this study was not taken into account. According to Chen (2013a:435), genre can have an impact on PV frequency, and this aspect could, therefore, be a valuable adjunct in an investigation into PV use.

Earlier in this chapter, it was suggested that South African L2 students differ from L2 students in other research because of the supposed continual exposure of South African L2 students to English. Yet the accuracy of this supposition needs to be verified by a closer inspection of students' backgrounds. Anecdotal evidence suggests that there might be students, especially those from rural areas and schools, for whom there is little such L2 exposure. Might such students be more likely to display the PV avoidance behaviour observed in countries where English is a foreign language? PV use by these students would provide an interesting contrast to that of students with high L1 exposure from a young age, and could inform remedial pedagogical practices. In this regard, research conducted by Gardner and Davies (2007:37) might prove relevant. They found the two everyday activities of reading and social networking to be better predictors of "formulaic language knowledge" (such as PV use) than

immersion in the target language. While their research was conducted with Chilean learners of English, this might be relevant information for rurally-based South African students.

Many researchers point out the importance of having a grammatical structure similar to that of a PV in the mother tongue for ease of PV attainment (Dagut & Laufer, 1985; Laufer & Eliasson, 1993). Except for Afrikaans (where PVs are referred to as *skeibare werkwoorde*), information about PV or similar language structures in the other nine South African languages was not readily available, and, therefore, no conclusions could be drawn as to this influence on PV acquisition. (Because of the fact that these nine languages are non-Germanic, it is highly likely that the PV structure is absent. In this regard, it should be noted that Afrikaans is a Germanic language.) The role of the mother tongue in the acquisition of problematic grammatical structures would be a worthwhile study, especially given the complexity of the South African context, and might inform problems with the acquisition of English on a broader level as well.

A further research focus that might be of interest pertains to the assumptions by most researchers in this field that L2 students prefer one-word alternatives over PVs. Whereas it is widely acknowledged that PVs present difficulties for novice L2 learners, the same reason has been given for avoidance by more advanced L2 learners. For example, Siyanova and Schmitt (2007:132) found that "length of time in a native English-speaking environment had no discernible effect on the likelihood of using multi-word verbs". In such scenarios, researchers are often inclined to rely on speculation and assumption to explain the phenomenon. On the other hand, interviewing L2 students would help to determine whether their preference for ALTs was the result of uncertainty, or of a conscious preference for a more formal style. (The difference in PV use between British and American native speakers, as reported on by Chen (2013a), comes to mind here.)

An aspect that was outside the scope of the current study but which would provide valuable insight is the identification of PVs that might be considered South African, having been coined in this country. To be most effective, the sourcing of such PVs would require a search of not only written material,

such as newspaper and magazine articles, advertisements, and textbooks, but also the spoken word. South African tendencies in word formation would be a valuable adjunct to research of this kind.

Finally, it should be kept in mind that this research study investigated PV use in student academic writing. Should it be possible to do a similar investigation using a speech corpus, differences between student speech and academic writing would deliver interesting and worthwhile results, as would the differences between L1 and L2 PV use in a spoken context.

Having considered possible future research topics concerning PV use, the chapter can be concluded. The next section will summarise the aspects discussed, and end with final comments on the research that was conducted in this study.

5.7 Conclusion

In this final chapter, the aims and research questions that drove this research study were reviewed. The research findings, as well as their broader implications, were discussed. It was then possible to assess the contribution that this research study might make to the South African educational environment, and the pedagogical implications that this might have. Thereafter, the possible limitations of the research were examined, and suggestions were made for future research topics.

In this research, the aim was to investigate the use of PVs in a South African environment, and, in particular, the use of PVs by L2 students. In order to obtain as complete a picture of PV use as possible, it was also necessary to investigate L1 student PV use. A corpus-based approach was used as the academic student writing of three institutions was available in electronic form, providing real-world data for the research. The data of the three institutions were separated into L1 and L2 subcorpora as required by the research aims, and software was used to extract PVs from the various subcorpora. This was first done at institutional level, allowing for the comparison of L1 and L2 PV frequencies within each institution, after which L1 and L2 PV frequencies were compared across institutions. The overall

L1 and L2 PV frequencies were then compared, which made it possible to address the research questions of the study.

In contrast to what was expected, the L2 students were found to use more PVs than the L1 students, and, therefore, did not display avoidance behaviour. However, a thorough investigation of the results further suggested that the L2 students did not use as varied a selection of PVs as did the L1 students, and that they were more prone to make errors in the use of PVs, compared to the L1 students. Both groups of students displayed a preference for ALTs over PVs in their writing, which, again, contradicted the expectation that the L2 students would be found to prefer the use of ALTs more often than the L1 students. In fact, the L1 students were found to use ALTs far more often that the L2 students. A consideration of the overall results suggested that L2 students, while not avoiding PVs, were more likely than the L1 students to use PVs on a once-off basis, using fewer ALTs than the L1 students, and making more mistakes than the L1 students when using PVs. On the other hand, the L1 student PV use displayed signs of having a more sophisticated vocabulary than the L2 students, and being more competent in their PV use. However, it was interesting to note that the two groups of students used many of the same high-frequency PVs.

There does not seem to have been great interest in the use of the PV in South African education. This might be as a result of a world-wide tendency to see the PV as primarily a feature of speech, and, thus, unsuitable in academic writing. Consequently, it would be seen as a waste of valuable teaching time to include it in the curriculum. Yet it is hoped that this research will have provided sufficient reasons for its inclusion in the classroom. Firstly, it has been shown that PVs are, in fact, suitable for inclusion in academic and other types of formal writing, as has been the case over many centuries. In this regard, it is necessary to distinguish between those PVs that are appropriate for use in formal writing, and those that fall into the category of slang. This distinction should form an important part of PV instruction. Secondly, even if the decision were to be made by the South African educational authorities to avoid the inclusion of PVs in academic writing entirely (and there are certainly

researchers who take this point of view (see §2.5)), learners would still need to be informed about the structure that they need to avoid. The third point refers to the aims of L2 education in South Africa. Is the aim simply to teach learners to write in English, or to become proficient in English overall? If the latter is true, then learners will certainly need to be instructed in the use of the PV, as, without its inclusion in one's speech, one cannot be thought of as having obtained a natural degree of fluency in English.

List of references

Alangari, M., Jaworska, S. & Laws, J. 2020. Who's afraid of phrasal verbs? The use of phrasal verbs in expert academic writing in the discipline of linguistics. *Journal of English for Academic Purposes,* 43(100814):1-13.

Aldukhayel, D.M. 2014. *The L2 exposure effect on avoidance of phrasal verbs by Arab ESL learners* Masters thesis. Colorado State University.

Alhatmi, S. 2023. Weighing up the effect of contextual cues in learning English phrasal verbs: Is context the answer to avoidance? *Journal of English Language Teaching and Applied Linguistics,* 5(3):1-14.

Atmowardoyo, H. 2018. Research methods in TEFL studies: Descriptive research, case study, error analysis, and R & D. *Journal of Language Teaching and Research,* 9(1):197-204.

Badem, N. & Simsek, T. 2021. A comparative corpus-based study on the use of phrasal verbs by Turkish EFL learners and L1 English speakers. *Advances in Language and Literary Studies,* 12(6):55-66.

Baker, P. 2006. *Using corpora in discourse analysis.* London: Bloomsbury.

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

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman grammar of spoken and written English.* London: Longman

Biber, D. & Reppen, R. 2015. *The Cambridge handbook of English corpus linguistics.* Cambridge University Press.

Blais, M.-J. 2012. *Explicit and implicit semantic processing of verb-particle constructions in L2.* PhD thesis. McGill University Libraries. (Unpublished).

Bolinger, D. 1971. *The phrasal verb in English.* Cambridge, Mass.: Harvard University Press.

Botha, Y.V. 2012. *Specification in the English nominal group with reference to student writing.* Doctor of Philosophy in Linguistics and Literary Theory. Potchefstroom Campus of the North-West University.

Boughey, C. 1998. Language and" disadvantage" in South African institutions of higher education: Implications of critical challenges to second language acquisition discourses for academic development practitioners. *South African Journal of Higher Education,* 12(1):166-173.

Brezina, V. 2018. *Statistics in corpus linguistics: A practical guide.* Cambridge University Press.

Bronshteyn, K.C. & Gustafson, T. 2015. The acquisition of phrasal verbs in L2 English: A literature review. *Linguistic Portfolios,* 4(1):91-99.

Bybee, J.L. 2006. From usage to grammar: The mind's response to repetition. *Language,* 82(4):711-733.

Cambridge international dictionary of phrasal verbs. 1997. Cambridge, UK: Cambridge University Press.

Cambridge Phrasal Verbs Dictionary, 2nd ed. 2006. Cambridge, United Kingdom: Cambridge University Press.

Carstens, A. 2016. Translanguaging as a vehicle for L2 acquisition and L1 development: students' perceptions. *Language Matters,* 47(2):203-222.

Carstens, A. & Eiselen, R. 2019. Designing a South African multilingual learner corpus of academic texts (SAMuLCAT). *Language Matters,* 50(1):64-83.

Celce-Murcia, M. & Larsen-Freeman, D. 1983. *The grammar book: An ESL/EFL teacher's course.* Boston, Mass.: Heinle & Heinle.

Chapelle. 2007. Technology and second language acquisition. *Annual Review of Applied Linguistics,* 27.

Chen, M. 2013a. Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students. *International Journal of Corpus Linguistics,* 18(3):418-442.

Chen, M. 2013b. Phrasal verbs in a longitudinal learner corpus: Quantitative findings. In: Granger, S., Gilquin, G. & Meunier, F. (eds.). *Twenty years of learner corpus research: Looking back, moving ahead.* Louvain-la-Neuve: Presses Universitaires de Louvain:89-101.

Chu, Y.Y. 1996. *Phrasal verbs for ESL students in Taiwan.* PhD thesis. The University of Texas at Arlington. (Unpublished).

Civan, A. & Coskun, A. 2016. The effect of the medium of instruction language on the academic success of university students. *Educational Sciences: Theory and Practice,* 16(6):1981-2004.

Collins COBUILD Phrasal Verbs Dictionary, 3rd ed. 2013. Great Britian: HarperCollins Publishers Limited.

Collins Online English Dictionary. 2021. [Online]. Available: https://www.collinsdictionary.com/dictionary/english [Accessed 10 September 2021].

Cook, V.J. 2010. The relationship between first and second language acquisition revisited. *The Continuum companion to second language acquisition*:137-157.

Cooper, P.A. 2016. *Academic vocabulary and lexical bundles in the writing of undergraduate psychology students.* PhD thesis (Unpublished). University of South Africa.

Cooper, T. 2017. Students' use of academic vocabulary in comparison to that of published writers: A corpus-driven analysis. *Stellenbosch Papers in Linguistics,* 47:133-152.

Creswell, J.W. 2009. *Research Design Qualitative, Quantitative, And Mixed Methods Approaches 3rd ed.* Thousand Oaks, California: Sage Publications Inc.

Dagut, M. & Laufer, B. 1985. Avoidance of phrasal verbs—A case for contrastive analysis. *Studies in second language acquisition,* 7(1):73-79.

Darwin, C.M. & Gray, L.S. 1999. Going after the phrasal verb: An alternative approach to classification. *Tesol Quarterly,* 33(1):65-83.

Department of Basic Education. 2021. Learning and teaching support materials. [Online]. Available: https://www.education.gov.za/Curriculum/LearningandTeachingSupportMaterials(LTSM).aspx#:~:text=Support%20Materials%20(LTSM)-,Learning%20and%20Teaching%20Support%20materials,-National%20Catalogues [Accessed 25 March 2024].

Dixon, R.M. 1982. The grammar of English phrasal verbs. *Australian Journal of Linguistics,* 2(1982):1-42.

Elenbaas, M.B. 2007. *The synchronic and diachronic syntax of the English verb-particle combination.* Doctoral dissertation. Utrecht: LOT.

The Free Dictionary. 2024. [Online]. Available: https://idioms.thefreedictionary.com/ [Accessed 22 August 2023].

Furey, E. 2024. *Descriptive Statistics Calculator*. [Online]. Available: https://www.calculatorsoup.com/calculators/statistics/descriptivestatistics.php.

Gardner, D. & Davies, M. 2007. Pointing out frequent phrasal verbs: A corpus-based analysis. *Tesol Quarterly,* 41(2):339-359.

Garnier, M. & Schmitt, N. 2015. The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research,* 19(6):645-666.

Garnier, M. & Schmitt, N. 2016. Picking up polysemous phrasal verbs: How many do learners know and what facilitates this knowledge? *System,* 59:29-44.

Gaston, M. 2004. *Avoidance of phrasal verbs by Spanish-speaking learners of English.* Masters thesis. California State University, Dominguez Hills. (Unpublished).

Gilquin, G. 2015. The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory,* 11(1):51-88.

Glosbe Multiligual Online Dictionary. 2024. [Online]. Available: https://glosbe.com/en/en/to%20carry%20across [Accessed 22 August 2023].

Gordon, S. & Harvey, J. 2019. *South Africans prefer their children to be taught in English.* [Online]. Available: https://theconversation.com/south-africans-prefer-their-children-to-be-taught-in-english-124304 [Accessed 5 October 2024].

Gorlach, M. 2008. Resultativeness: Constructions with phrasal verbs in focus. *Between grammar and lexicon*:255-290.

Haider, A.S., Saed, H.A., Hussein, R.F., Al-Abbas, L.S. & Meqdadi, S.R. 2020. The impact of English proficiency on university students' use of one-word or phrasal verbs. *International Journal of Innovation, Creativity and Change,* 14(5):1184-1196.

Haugh, S. & Takeuchi, O. 2023. Learner knowledge of English phrasal verbs: Awareness, confidence, and learning experiences. *International Journal of Applied Linguistics*.

He, X. 2017. *Second language acquisition of particle-verb constructions in English by adult Mandarin speakers.* Masters thesis. University of North Carolina.

Henrich, J., Heine, S.J. & Norenzayan, A. 2010. Most people are not WEIRD. *Nature,* 466(7302):29-29.

Huddleston, R.D., Pullum, G.K. & Bauer, L. 2002. *The Cambridge Grammar of the English Language.* Cambridge University Press.

Hulstijn, J.H. & Marchena, E. 1989. Avoidance: Grammatical or semantic causes? *Studies in second language acquisition,* 11(3):241-255.

Hummel, K.M. 2014. *Introducing Second Language Acquisition Perspectives and Practices*. Newark: John Wiley & Sons, Incorporated. [Online]. Available: https://public.ebookcentral.proquest.com/choice/PublicFullRecord.aspx?p=7103833.

Immelman, S. & Cooper, T. 2023. Come on, carry on: Phrasal verb use in undergraduate writing at a South African university. *Journal for Language Teaching,* 57(1):1-26.

Internet Archive. 2014. [Online]. Available: archive.org [Accessed 12 March 2021].

Kamarudin, R. 2014. *A study on the use of phrasal verbs by Malaysian learners of English.* PhD thesis. University of Birmingham. (Unpublished).

Kennedy, A.G. 1920. *The modern English verb-adverb combination*. California: Stanford University. [Online]. Available: https://archive.org/details/modernenglishve00kenngoog/page/n2/mode/2up [Accessed 12 October 2022].

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. 2004. *SketchEngine*. [Online]. Available: https://www.sketchengine.eu/.

Kleinmann, H.H. 1977. Avoidance behavior in adult second language acquisition 1. *Language learning,* 27(1):93-107.

Kovács, É. 2002. Properties of Verbs Which Constitute Phrasal Verbs. *Eger Journal of English Studies,* 3:109-128.

Laufer, B. 2000. Avoidance of idioms in a second language: The effect of L1-L2 degree of similarity. *Studia linguistica,* 54(2):186-196.

Laufer, B. & Eliasson, S. 1993. What causes avoidance in L2 learning: L1-L2 difference, L1-L2 similarity, or L2 complexity? *Studies in second language acquisition,* 15(1):35-48.

Liang, A. 2022. *China Covid: Area around world's biggest iPhone plant locked down.* [Online]. Available: https://www.bbc.co.uk/usingthebbc/terms/can-i-use-bbc-content/ [Accessed 13 March 2024].

Liao, Y. & Fukuya, Y.J. 2004. Avoidance of phrasal verbs: The case of Chinese learners of English. *Language learning,* 54(2):193-226.

Linguee. 2024. [Online]. Available: https://www.linguee.com/english-spanish/translation/crawl+away.html [Accessed 10 November 2024].

Liu, D. 2011. The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *Tesol Quarterly,* 45(4):661-688.

Liu, D. & Myers, D. 2020. The most-common phrasal verbs with their key meanings for spoken and academic written English: A corpus analysis. *Language Teaching Research,* 24(3):403-424.

Lu, Z. & Sun, J. 2017. Presenting English polysemous phrasal verbs with two metaphor-based cognitive methods to Chinese EFL learners. *System*, 69. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0346251X17306589 [Accessed 14 March 2024].

Ludwig.guru. 2024. *Write in English at your best.* [Online]. Available: https://ludwig.guru/s/to+send+this+ [Accessed 22 August 2023].

Lutrin, B. & Pincus, M. 2007. English handbook and study guide. *Birnam Park: Belut Books*.

Macaro, E. 2010. Second language acquisition: the landscape, the scholarship and the reader. *In:* Macaro, E. (ed.) *Continuum companion to second language acquisition.* Bloomsbury.

Mahpeykar, N. & Tyler, A. 2015. A principled cognitive linguistics account of English phrasal verbs with up and out. *Language and Cognition,* 7(1):1-35.

Mathew, P., Nesi, H. & Vincent, B. 2019. Corpus from scratch: Collecting and processing a sizeable EAP corpus in a (relatively) resource-poor context. BALEAP Biannual Conference.

Mazaherylaghab, H. 2015. *Iranian learner English: A corpus-based study of phrasal verb usage.* Doctoral dissertation. Albert-Ludwigs-Universität.

McArthur, T. 1989. The long-neglected phrasal verb. *English Today,* 5(2):38-44.

McEnery, T. & Hardie, A. 2012. *Corpus linguistics: Method, theory and practice.* New York: Cambridge 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.* Abingdon: Routledge.

McPartland-Fairman, P. 1989. *The processing of phrasal verbs by native and nonnative speakers of English.* Doctoral dissertation. New York: City University of New York.

Merriam-Webster Online Dictionary. 2024. Merriam-Webster, Incorporated. [Online]. Available: https://www.merriam-webster.com/dictionary/go%20back/ [Accessed 30 May 2023].

Morales, A.E. 2000a. *Use and comprehension of English phrasal verbs among native Spanish speakers.* Doctoral dissertation. University of Kansas.

Morales, A.E. 2000b. *Use and comprehension of English phrasal verbs among native Spanish speakers.* Doctoral dissertation. University of Kansas.

Munje, P.N. & Jita, T. 2020. The impact of the lack of ICT resources on teaching and learning in selected South African primary schools. *International Journal of Learning, Teaching and Educational Research,* 19(7):263-279.

Murphy, V.A. 2010. The relationship between age of learning and type of linguistic exposure in children learning a second language. In: Macaro, E. (ed.). *The Continuum companion to second language acquisition.* London, New York: Bloomsbury:158-178.

Myers, D.J. 2018. *Analyzing semantic/usage distributions of phrasal verbs across registers: A corpus-based study on PVs' most frequent senses in academic and spoken English.* Doctoral dissertation. University of Alabama

Oxford English Dictionary. 2023. Oxford University. [Online]. Available: https://www.oed.com/discover/introduction-to-south-african-english?tl=true.

Oxford Learner's Dictionaries. 2024. Oxford University Press. [Online]. Available: https://www.oxfordlearnersdictionaries.com/definition/english/act_2#act_idmg_2 [Accessed 23 April 2023].

Özkayran, A. & Yilmaz, E. 2020. Analysis of higher education students' errors in English writing tasks. *Advances in Language and Literary Studies,* 11(2):48-58.

Qiu, C.W. 2018. *A comparison of phrasal verbs in Singapore English with British and American English.* Thesis (M.A.) National Institute of Education, Nanyang Technological University.

Rayson, P. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics,* 13(4):519-549.

Reverso. 2024. *ReversoDictionary.* [Online]. Available: https://dictionary.reverso.net/english-definition/ [Accessed 22 August 2023].

Riguel, E. 2014. Phrasal verbs: usage and acquisition. *Athens Journal of Philology,* 1(2):111-126.

Roohani, A. & Heidari Vincheh, M. 2023. Effect of game-based, social media, and classroom-based instruction on the learning of phrasal verbs. *Computer Assisted Language Learning,* 36(3):375-399.

Saraswati, M. 2024. *The correlation between students' habit in watching English movies and their mastery of phrasal verbs (A study at class 2020 English education students of Raden Intan State Islamic University Lampung).* S1 degree. Lampung: Raden Intan State Islamic University.

Saville-Troike, M. & Barto, K. 2017. *Introducing second language acquisition.* Cambridge University Press.

Sawyer, J.H. 2000. Comments on Clayton M. Darwin and Loretta S. Gray's "Going after the phrasal verb: An alternative approach to classification". A reader reacts. *Tesol Quarterly,* 34(1):151-159.

Schachter, J. 1974. An error in error analysis. *Language learning,* 24(2):205-214.

Schmitt, N. & Redwood, S. 2011. Learner knowledge of phrasal verbs: A corpus-informed study. In: Meunier, F., De Cock, S., Gilquin, G. & Paquot, M. (eds.). *A taste for corpora: In honour of Sylviane Granger.* Amsterdam: John Benjamins Publishing Company:173-209.

Scott, L. 2015. *English lingua franca in the South African tertiary classroom: recognising the value of diversity.* Masters degree. Stellenbosch: Stellenbosch University.

Scott, M. 2022. *WordSmith Tools version 8 (64 bit version)*. [Online]. Available.

Siyanova, A. & Schmitt, N. 2007. Native and nonnative use of multi-word vs. one-word verbs. *IRAL,* 45(2007):119-139.

Southwood, F., White, M.J., Brookes, H., Pascoe, M., Ndhambi, M., Yalala, S., Mahura, O., Mössmer, M., Oosthuizen, H. & Brink, N. 2021. Sociocultural factors affecting vocabulary development in young South African children. *Frontiers in Psychology,* 12:642315.

Spada, N. & Lightbown, P.M. 2019. Second language acquisition. In: Schmitt, N. & Rodgers, M.P.H. (eds.). *An introduction to applied linguistics.* 3rd ed. London and New York: Routledge:111-127.

StackExchange. 2024. *English language & usage.* [Online]. Available: https://english.stackexchange.com/questions/319775/ [Accessed 22 August 2023].

Statistics South Africa. 2024. [Online]. Available: https://census.statssa.gov.za/assets/documents/2022/Census_2022_SG_Presentation_10102023.pdf [Accessed 17 July 2024].

Thesaurus.plus. 2024. [Online]. Available: https://thesaurus.plus/synonyms/crawl_away [Accessed 22 November 2023].

Thim, S. 2012. *Phrasal verbs: The English verb-particle construction and its history.* Berlin/Boston: Walter de Gruyter.

Times Higher Education Sub-Saharan Africa University Rankings 2024. [Online]. Available: https://www.timeshighereducation.com/student/best-universities/best-universities-africa [Accessed 14 March 2024].

Trebits, A. 2009. The most frequent phrasal verbs in English language EU documents–A corpus-based analysis and its implications. *System,* 37(3):470-481.

Van Rooy, B. & Coetzee-Van Rooy, S. 2015. The language issue and academic performance at a South African University. *Southern African Linguistics and Applied Language Studies,* 33(1):31-46.

Van Rooy, B. & Kruger, H. 2015. The case for an emergentist approach. *Stellenbosch Papers in Linguistics Plus,* 48:41-67.

Van Rooy, B. & Terblanche, L. 2006. A corpus-based analysis of involved aspects of student writing. *Language Matters: Studies in the Languages of Southern Africa,* 37(2):160-182.

Waibel, B. 2007. *Phrasal verbs in learner English: A corpus-based study of German and Italian students.* Freiburg (Breisgau) University.

Wei, Y. 2021. Use of English phrasal verbs of Chinese students across proficiency levels: A corpus-based analysis. *International Journal of TESOL Studies,* 3(4):25-41.

Weisser, M. 2016. *Practical corpus linguistics: An introduction to corpus-based language analysis.* John Wiley & Sons.

White, B.J. 2012. A conceptual approach to the instruction of phrasal verbs. *The Modern Language Journal,* 96(3):419-438.

Wiktionary. 2024. [Online]. Available: https://en.wiktionary.org/wiki/joke_around; https://www.yourdictionary.com/joke-around [Accessed 10 November 2023].

Wilcoxon, E.M. 2014. *A corpus-based study of the use of prepositional verbs in second language emergent academic writing.* Master of Arts. The University of Texas at El Paso.

Wild, C.K. 2010. *Attitudes towards English usage in the late modern period: the case of phrasal verbs.* University of Glasgow.

Wissing, D. 2002. Black South African English: a new English? Observations from a phonetic viewpoint. *World Englishes,* 21(1):129-144.

Word Hippo. 2008. [Online]. Available: https://www.wordhippo.com/what-is/another-word-for/carry_across.html [Accessed 22 August 2023].

Wordnik. 2024. [Online]. Available: https://www.wordnik.com/words/joke%20around [Accessed 15 November 2023].

WordReference.com. 2024. *WordReference.com Language Forums.* [Online]. Available: https://forum.wordreference.com/threads/send-out-send-through-and-send-over.1338005/ [Accessed 22 August 2023].

Yasuda, S. 2010. Learning phrasal verbs through conceptual metaphors: A case of Japanese EFL learners. *Tesol Quarterly,* 44(2):250-273.

You, Y.-S. 1999. Avoidance phenomena of phrasal verbs by Korean learners of English. *English teaching,* 54(3):135-155.

YourDictionary. 2024. [Online]. Available: https://www.yourdictionary.com/joke-around [Accessed 10 November 2023].

Zarifi, A. & Mukundan, J. 2014. Creativity and unnaturalness in the use of phrasal verbs in ESL learner language. *Southeast Asian Journal of English Language Studies,* 20(3):51-62.

Zhang, X. & Ju, W. 2019. Exploring multiple constraints on second language development of English polysemous phrasal verbs. *Applied Psycholinguistics,* 40(5):1073-1101.

Zhou, S. & Wang, H. 2024. "Let's move on to the Recommendations." The use of phrasal verbs in business presentations of university students in Hong Kong. *STiLE-Scholarship of Teaching in Language Education,* 2(1):1-28.

Appendix 1

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Humanities
Fakulteit Geesteswetenskappe
Lefapha la Bomotho

Humanities 100
Since 1919

18 May 2022

Dear Mrs S Immelman

Project Title:	Phrasal verbs in undergraduate writing: A focus on South African second-language students
Researcher:	Mrs S Immelman
Supervisor(s):	Dr PA Cooper
Department:	Afrikaans
Reference number:	04181956 (HUM020/0222)
Degree:	Doctoral

Thank you for the application that was submitted for ethical consideration.

The Research Ethics Committee notes that this is a literature-based study and no human subjects are involved. The application has been **approved** on 18 May 2022 with the assumption that the document(s) are in the public domain. Data collection may therefore commence, along these guidelines.

Please note that this approval is based on the assumption that the research will be carried out along the lines laid out in the proposal. However, should the actual research depart significantly from the proposed research, a new research proposal and application for ethical clearance will have to be submitted for approval.

We wish you success with the project.

Sincerely,

Prof Karen Harris
Chair: Research Ethics Committee
Faculty of Humanities
UNIVERSITY OF PRETORIA
e-mail: tracey.andrew@up.ac.za

Appendix 2

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

SCHOOL OF HUMAN AND COMMUNITY DEVELOPMENT
PSYCHOLOGY

03 May 2022

University of Pretoria Research Ethics Committee

To whom it may concern,

Re: Permission to use data corpus for PhD study: Susan Immelman

Susan Immelman would like to conduct a study based on a corpus of data collected by her supervisor Dr Trish Cooper from Psychology students at the University of the Witwatersrand.

This letter serves to confirm that Dr Cooper's study was approved by the Wits Research Ethics Committee and permission was also given from the Head of the Department of Psychology at the time, Prof Andrew Thatcher.

This letter also gives permission from myself, the current Head of the Department of Psychology for this data to be used by Susan Immelman. I cannot grant permission from the Wits Ethics Committee. My permission is also contingent on ethics clearance being obtained by Susan Immelman for her study.

Should you have any queries, please do not hesitate to contact me.

Yours sincerely,

Prof Carol Long, Head of Department, Psychology
☎ +2711 717 4510 ✉ carol.long@wits.ac.za

Umthombo Building – 2nd Floor, Room U211, Braamfontein East Campus, Private Bag 3, WITS 2050
T +27 11 717 4503/89 | E psych.SHCD@wits.ac.za | www.wits.ac.za/shcd/psychology/

Re: Toestemmingsbrief

External | Ethical clearance ×

Tobie Van Dyk <Tobie.VanDyk@nwu.ac.za>
to me, DJ, Juan, trish.cooper

Mon, 28 Feb 2022, 07:54

Translate to English

Dear Sue

Thank you for your email. I had a look at the outline of your study and am satisfied that you can go ahead. From my side, I therefore give permission for use of the SAMULKCAT corpus for your research. You will have to engage Juan Steyn and colleagues at SADiLaR on how to access the corpus and do searches. Please note that they can't "clean" the corpus for your. You will have to indicate to them exactly what it is that you wish to do and then they will guide you as to where to find it on the corpus platform, of available.

Kind regards
Tobie

Prof. Tobie van Dyk
Skool vir Tale | School of Languages | Sekolo sa Dipuo