Stellenbosch Papers in Linguistics, Vol. 47, 2017, 89-114
doi: 10.5842/47-0-263

South African students' use of delexical multiword units: The trouble with high-frequency verbs

Ruth Scheepers

Department of English Studies, University of South Africa, South Africa
E-mail: scheera@unisa.ac.za

Abstract
This article describes a corpus-linguistic investigation of undergraduates' production of delexical multiword units (MWUs) comprising high-frequency verb + noun combinations. The aim was to shed more light on the difficulties these deceptively simple combinations pose for writers in a multilingual South African context. Two corpora of learner writing from different areas of English studies (literature and communication for law) and a reference corpus of scholarly writing were compared, focusing on the frequency of MWUs in the student corpora and errors in these combinations. That these MWUs and the common verbs they feature are "error-prone" (Altenberg and Granger 2001:179) in learner language is well attested in current research. This study found that student writers did indeed have difficulty producing error-free delexical MWUs. A detailed analysis of their errors found that these were caused mainly by the verb in the combination, particularly verb collocation. These findings highlight the difficulties these combinations pose for South African learners. Such combinations are common in everyday language and academic writing, and the findings underline the importance of a sound knowledge of high-frequency verbs and their collocations for students writing in an academic milieu.

Keywords: delexical multiword units, high-frequency verbs, South African student writing, error analysis

1. Introduction

Research in recent decades has seen growing consensus on the formulaic nature of language, and the view that a great deal of text is made up of "non-arbitrary and non-random phrases and patterns" (Kaszubski 2000:2) is generally accepted by scholars. Corpus linguistic research has been particularly productive in showing that "language is made up of not only individual words, but also a great deal of formulaic language" (Martinez and Schmitt 2012:299), and a great deal of research has been devoted to phraseology and to explaining various lexical patterns. It is generally accepted that formulaic language is vital to the native-like and idiomatic production of language (Erman and Warren 2000; Granger 1998; Kaszubski 2000; Nesselhauf 2003; Wray 2002). Wray (2002:13), for instance, believes that formulaicity is "all-pervasive in language data", and that words belong with other words at the most basic level, what Sinclair (1991:110) refers to as "unrandomness". Sinclair assumed that, rather than using isolated words in rule-governed sequences, speakers tend to use ready-made linguistic forms (Léon 2007). Many researchers (Erman and Warren 2000; Sinclair 1991; Wray 2002) believe that neither an analytical nor a holistic process alone can "accommodate both the linguistic competence of the ideal native speaker and listener and the idiomatic choice of one grammatical string over another" (Wray 2002:15). Learners who are proficient in a language understand that certain words typically occur in a particular structure, such as "SOMETHING UNDESIRABLE is/are rife in LOCATION/TIME" (Schmitt and Carter 2004:8). This is the notion of idiomaticity – a sense of the "salience", as Granger (1998) puts it, and an awareness of the "conventionality and naturalness of some expressions" (Kaszubski 2000:1).

While many different types of word combinations are referred to under the umbrella term "formulaic language", the focus in this article is on collocations made up of a chunk of language featuring a seemingly simple high-frequency verb and an eventive noun[1], as in *take a walk*. Nesselhauf's (2003) definition of collocations informs this study: the term "collocation" is used "in a phraseological rather than in a frequency-based sense" (2003:224), indicating a particular combination of words rather than words that co-occur in a particular span (cf. Sinclair 1991). Nizonkiza and Van de Poel (2014:302), who provide a detailed discussion of collocations, note that this approach is "characterised by the varying degrees of fixedness and substitutability of collocations". Nesselhauf distinguishes between free combinations (Howarth 1998a), where possible substitution relies on "semantic properties", and collocations, in which restriction on substitution is somewhat "arbitrary" (Nesselhauf 2003:225). This notion of "restricted sense" (Nesselhauf 2003:225) is at the core of her definition. The collocations in this study lie somewhere on the continuum between free combinations and idioms, thus falling into Howarth's (1998a, 1998b) category of restricted (or semi-restricted) collocations.

Table 1: Howarth's (1998a:28) collocational continuum

	Free combinations	**Restricted collocations**	**Figurative idioms**	**Pure idioms**
Lexical composites Verb + noun	*Blow a trumpet*	*Blow a fuse*	*Blow your own trumpet*	*Blow the gaff*
Grammatical composites Preposition + noun	*Under the table*	*Under attack*	*Under the microscope*	*Under the weather*

Stubbs (2001) believes that although these combinations may not be idioms, they are idiomatic in that they are used in a particular way by native speakers. They differ thus from free combinations; although most allow for some replacements, these are sometimes seemingly arbitrary. They combine with a noun phrase to form "relatively idiomatic expressions", which "form a cline of idiomaticity" between expressions that are clearly idiomatic, such as *have a look* and *make a killing*; and those that "retain the core meaning of these verbs", as in "*we have an extra one, he made a sandwich*" (Biber et al. 1999:1026). Between these two types of expressions are a host of relatively idiomatic phrases "such as *have a chance, take a walk*":

[1] An eventive noun is the noun in the combination that carries the major part of the meaning.

although the core meaning of the individual words is retained, these types of expression tend to take on a more idiomatic meaning (Biber et al. 1999:1027), and can mostly be replaced by a single verb (Stubbs 2001). Learners' production of such combinations, I argue, reflects one specific set of indicators of the depth of their vocabulary knowledge.

Scholars such as Howarth (1998a, 1998b) and Nesselhauf (2005) have pointed to the difficulties learners have in mastering formulaic expressions such as the restricted or semi-restricted combinations or collocations that are the focus of this article. Wray (2002:ix) supports this, noting that native speakers (NSs) tend to use formulaic language as "an easy option in their processing and/or communication". In the early stages of both first and second language acquisition, learners rely a great deal on formulaic language. However, paradoxically, formulaic language seems to be intermediate and advanced L2 learners' biggest obstacle in achieving native-like fluency, "because the learner lacks the necessary sensitivity and experience that will lead him or her unerringly away from all the grammatical ways of expressing a particular idea except the most idiomatic" (Wray 2000:463). Such formulaic expressions are often difficult for learners to understand, even when NSs would regard them as fairly transparent (Martinez and Schmitt 2012) or as more transparent than idioms. However, this transparency does not mean that these expressions are always compositional (Erman and Warren 2000:54). They are not difficult for learners to decode (read and understand), but they cause difficulties when it comes to encoding them, because the learner simply has to know the conventional way of saying these things (Farghal and Obiedat 1995; Martinez and Schmitt 2012; Nesselhauf 2003). Such combinations occur frequently in academic discourse; Shirato and Stapleton (2007) claim, for instance, that many high-frequency clusters occur with greater frequency than some common single words and pose great difficulties for ESL learners, making them particularly important for learners of English in higher education contexts. In response to this increasing awareness that "language is made up of not only individual words, but also a great deal of formulaic language" (Martinez and Schmitt 2012:299), Shin and Nation (2008), Simpson-Vlach and Ellis (2010) and Martinez and Schmitt (2012) have all compiled lists of MWUs which they believe should form part of teaching materials and practice.

Many collocations used in academic English writing feature high-frequency words (Algeo 1995; Altenberg and Granger 2001; Howarth 1998a, 1998b; Langer 2004). The three verbs selected for this study, *have*, *make* and *take*, were chosen because they are highly polysemous and have various language-specific tendencies resulting in specialised meanings, collocations and idiomatic uses (Viberg 1996; Altenberg and Granger 2001). Research has shown that these tend to be problematic for foreign and L2 learners (Altenberg and Granger 2001; Kaszubski 2000; Lee and Chen 2009; Wang and Shaw 2008; Yan 2006).

Such high-frequency verbs as *have*, *go*, *do*, *say*, *take*, *give*, *get* and *make* have been variously referred to as "light verbs" (Live 1973; Wittenberg and Piñango 2011), "small verbs", "support verbs" (Langer 2004; Ronan and Schneider 2015) and "delexical verbs" (Altenberg and Granger 2001; Howarth 1998a, 1998b). This is because they tend to occur in combinations where the noun carries the main semantic weight of the expression. In other words, their meaning is defined by the company they keep (Firth 1957, cited in Léon 2007), while they themselves carry little meaning (Howarth 1998a). The words with which they collocate are often not arbitrary but 'restricted' collocations (Howarth 1998a, 1998b) and are referred to in this study as "delexical MWUs". They are also termed "expanded predicates" (Algeo 1995), "stretched verb constructions" (Nesselhauf 2005), "support verb constructions" (Langer 2004), "periphrastic verbal constructions" (Wierzbicka 1982), "phrasal verb types" (Stein 1991) and "support" verb constructions (Langer 2004; Ronan and Schneider 2015). Algeo (1995:203-204), for instance, defines an "expanded predicate" as "an idiomatic verb-object construction in which the verb (e.g. *do*, *give*, *have*, *make* or *take*) is semantically general and the object is semantically specific (such as *somersault*, *nod*, *rest*, *promise*, or *walk*)". In these combinations, the noun is derivationally related to a verb that is roughly synonymous with the whole combination: "the meaning of *make an arrangement*, for example, largely corresponds to the meaning of *arrange*" (Nesselhauf 2005:20). As the verb carries very little meaning, or is semantically empty, it can be referred to as "delexical" (Nesselhauf 2004:109) or as "light" (Nesselhauf 2005:21).

Nesselhauf (2005:21) observes that "most restrictive definitions are those combinations of *make*, *take*, *give*, and *have* with an indefinite article and an eventive noun that is identical in form to the verb". This last definition is the one I use for the core delexical MWUs in this study, following Algeo's (1995) terminology. Such combinations are very common in speech and writing. For instance, Stubbs (2001:32-33) made a search for the lemma *take* in a corpus of over two million words. While he found 400 examples, *take* was used in its literal sense of "grasp with the hand" in only about 10% of these occurrences. By far the most common use of the verb was in delexical combinations such as *take a deep breath*, *take a photograph* and *take a decision* (Stubbs 2001:32), emphasising that delexicalisation is common in English.

In this study, Biber et al.'s (1999:1026-7) "relatively idiomatic" combinations are regarded as *core* delexical MWUs (Algeo 1995), where the single-word verb and the noun are identical, both in form and in meaning, and an indefinite article is present, e.g. *take a walk* = *walk*. If the form of the noun differs in any way, or if the article use is different, the combination is classed as a *pseudo* delexical MWU (Algeo 1995). According to this definition, the MWUs in this study are collocations made up of monotransitive verb patterns (verb + noun), which occur on Howarth's (1998a, 1998b) continuum, or on Erman and Warren's (2000) cline, somewhere between free combinations and idioms.

2. High-frequency verb-noun combinations and the difficulties they cause learners

Two seemingly contradictory observations about high-frequency verbs have been made. First, learners tend to overuse them (Altenberg and Granger 2001; Kaszubski 2000; Wang and Shaw 2008; Gilquin 2007; Lee and Chen 2009; Laufer and Waldman 2011). Hasselgren (1994:250) notes that "core words – learnt early, widely usable, and above all safe (*because* they do not show up as errors) are hugely overused, even among learners sufficiently advanced to have been weaned off them". However, learners also experience particular difficulty with these words and may therefore underuse them, especially when they are used in multiword combinations that are restricted or semi-restricted. Sinclair's (1991:79) "underuse hypothesis" posits that learners tend to avoid high-frequency verbs, especially where they are part of idiomatic phrases, making use instead of "larger, rarer or clumsier words which make their language sound stilted and awkward". Shirato and Stapleton (2007:409) arrived at similar findings in their study, which compared a small corpus of spoken language from adult Japanese learners of English to an established NS corpus. The Japanese learners differed particularly in their underuse of delexical verbs, using them relatively infrequently in their speech when compared to the NSs. The learners were seemingly reluctant to use multiword delexical clusters such as *get done* and *get locked in* and unaware that such clusters may be more appropriate in spoken language than more formal single-item verbs (Shirato and Stapleton 2007:407).

Altenberg and Granger (2001) used a computerised corpus of EFL writing comprising two samples from the International Corpus of Learner English (ICLE) database, one by advanced French-speaking learners of English and a second by Swedish learners, all in their second or third year of studying English at university. Focusing on the grammatical and lexical patterning of the verb *make* to establish whether learners overused or underused such high-frequency verbs and whether they were likely to cause errors (Altenberg and Granger 2001:178), they found that both language groups underused *make* in delexical structures, while conversely overusing the verb when it was used causatively (e.g. *to make somebody believe something*). Examining collocates of *make* that occurred at least twice in the corpora, they found that the learner corpora confirmed Sinclair's underuse hypothesis (Sinclair 1991:79), revealing that learners may simultaneously overuse a high-frequency verb and underuse its delexical structures. The non-native speakers (NNSs) also misused these delexical structures: this category accounted for most errors involving *make* in the corpus. Even at an advanced level of proficiency, EFL learners had difficulty with high-frequency verbs such as *make* (Altenberg and Granger 2001:189).

In a more recent study of advanced French learners' knowledge of *make* collocations, Gilquin (2007) found that learners tended to underuse collocations with *make*, unless they had a direct equivalent in their L1 and using it would be less likely to cause errors. On the other hand, Liu and Shaw (2001), who used a contrastive corpus analysis approach to investigate the use of the word *make* in two main corpora – the Chinese-speaking Learners of English (CSLE) and the Native Speakers of English (NSE) corpora – found that learners of English used *make* far more frequently than NSs, regardless of the category of text considered.

In their study, Lee and Chen (2009) used a multiple-comparison approach to investigate words and phrases that were overused and underused by learners. They compiled "three complementary types of corpora" (2009:151): the Chinese Academic Written English (CAWE) corpus; the Expert Journal Articles (EXJA) corpus; and a section of the British Academic Written English (BAWE) corpus, comprising high-scoring student essays by native English speakers, which they called the "sub-corpus BAWE-L". Using a keyword analysis, they identified positive keywords that were overused when compared to the reference corpus and those that were significantly less frequent or 'underused' negative keywords (Lee and Chen 2009:152). Like Altenberg and Granger (2001), they found that very high-frequency common words such as *make*, *besides*, *get* and *help* were among those most overused. However, further qualitative investigations of concordances and collocations found, in contrast to Altenberg and Granger (2001), that the high-frequency verb *make* was used much more frequently by Chinese learners in 'light verb' or 'delexical' constructions than by writers in the EXJA and BAWE-L corpora. However, many of these expressions were in fact unidiomatic or unnativelike; these so-called 'simple' words in their lists occurred in "recurrent problematic patterns rather than being randomly used" (Lee and Chen 2009:154). Wang and Shaw's (2008:203) comparison of the collocational errors of Swedish and Chinese university students learning English also revealed that advanced learners experienced difficulties with collocations formed with what they refer to as "frequent, high-utility dynamic verbs", reflecting a lack of collocational competence among learners when using verbs such as *have*, *do*, *make* and *take*.

Kaszubski (2000), investigating high-frequency verbs often used in delexical structures (in this case *be*, *have*, *make*, *take*, *do* and *get*) and their various combinations in corpora of writing by Polish, Spanish and French learners of English, found that delexical collocations were underused by these learners. Likewise, in their study of verb-noun collocations in the writing of three groups of Hebrew-speaking second-language learners at different proficiency levels, Laufer and Waldman (2011) found that NNSs used fewer collocations overall, using significantly fewer of these combinations than the NSs. As in Altenberg and Granger's (2001) study, Laufer and Waldman (2011:664) found that, compared to NSs, learners tended to *underuse* collocations containing what they refer to as "core" verbs (*be*, *have*, *make*), and that collocation posed difficulties even for advanced learners. Like other scholars mentioned in Section 1, Laufer and Waldman (2011:648) stress the importance of MWUs "as a necessary component of second-language lexical competence in addition to the knowledge of single words", and they underline the view that knowledge of MWUs improves both the quality and the fluency of language, spoken and written.

Nesselhauf (2005) made a study of English collocation use by German university students, including a detailed error analysis. Distinguishing stretched verb constructions (SVCs), her term for delexical MWUs, from collocations, she found that SVCs proved particularly difficult for language learners (Nesselhauf 2005:211). These combinations made up over 20% of all the collocations identified in her German learner corpus (Nesselhauf 2005:211). Almost a quarter of these were judged as unacceptable. When those that were deemed questionable were taken into consideration as well, 43% of all SVCs were judged deviant. She found that students were more inclined to make errors in combinations where the verb took a relatively wide range of nouns than where the number of nouns was more restricted (what she terms "RC1 restricted collocations"), such as *pay attention* and *take a picture*. RC1 combinations are "more often acquired and produced as wholes", whereas the less restricted collocations allow for more creative combinations, and thus more potential errors (Nesselhauf 2003:233). In fact, in her study, SVCs with light verbs, especially high-frequency ones, were not in the main produced incorrectly; but because learners used these structures often and because they often contained high-frequency verbs, they got them wrong in the "absolute sense" (Nesselhauf 2005:212). In other words, as their mistakes in using these high-frequency verbs were seen often, it was assumed that learners found them difficult.

Few studies have as yet dealt with MWUs in the South African context, and there is thus an opportunity for more research in this area. One scholar who has been at the forefront of research into collocations in South Africa is Deogratias Nizonkiza (North West University, South Africa). For instance, in a study addressing the difficulties collocations pose to students, Nizonkiza, Van Dyk and Louw (2013) used a productive collocation test based on Laufer and Nation (1999) to test the ability of a group of South African university students to produce collocates in sentences where the first two letters of the target word were provided. Taking 80% as mastery level for each word level, the authors found that only collocates from the 2000-word level had been mastered by all students, and that over 60% of students had not mastered the 3000-word level or the Academic Word List (AWL) (Coxhead 2000), regarded by Nation (1990, in Nizonkiza et al. 2013:166) to be the minimum level of productive vocabulary required for success at university level (Nizonkiza et al. 2013).

The studies discussed in this section have highlighted the difficulties these MWUs can pose for learners, regardless of their language background, and yet such units of language are regarded as an essential component of lexical competence, together with the knowledge of individual words (Laufer and Waldman 2011:648). In addition, these combinations are common in academic writing. Thus, they are particularly crucial for second-language learners to master, as they improve the fluency and quality of both spoken and written language (Nesselhauf 2004; Pawley and Syder 1983; Shirato and Stapleton 2007; Wray 2002; Kaszubski 2000; Shin and Nation 2008).

3. Methodology

This section discusses the participants of the study, the compilation of the corpora and the data analysis.

3.1 Participants

Participants[2] were undergraduate students at a South African open distance learning (ODL) university, enrolled in two English courses, one teaching first-level English literature and the other English communication for law. The sample comprised 298 students in all, 175 Literature and 123 Law students. For a full explanation of the sampling method, please refer to Scheepers (2014, 2016). These students included speakers of all 11 official languages in South Africa. They were all 'learners' in the sense that they could be regarded as novice writers in the context of student academic writing, but the group did include some speakers of English as a first language. Thus, they were not all 'learners' of the language in the sense that Granger (1998) or others apply the term (i.e. to second or foreign language learners).

3.2 Compilation of the corpora

The student corpus was compiled from essays written by these students under timed examination conditions. As participants were enrolled at a distance education institution, the only opportunity to elicit authentic writing from them was during examinations, when they had no access to reference material, the internet or other possible input. The content of the essays was determined by the requirements of the courses concerned – argumentative essays on a legal question in the Law module and an essay or short questions on a prescribed novel, poem or extract in the Literature module. Once these essays had been transcribed, they formed a corpus of 206,173 words (tokens), made up of writing by literature students (henceforth the *Student Lit* corpus, comprising 142,655 words) and law students (the *Student Law* corpus of 63,518 words).

The Expert corpus, used as a reference corpus, was made up of the study material written by experts in these disciplines (i.e. Literature and Communication for Law), which the students read during the semester. The complete Expert corpus contained 192,060 words in total: 144,231 words in the *Expert Lit* and 47,829 words in the *Expert Law* corpus[3].

[2] The article reports on one phase of a larger study (see Scheepers 2014).

[3] In this article, for reasons of space and because the focus is on a comparison of two student corpora, the results of the comparisons between the Student and Expert corpora are not reported in detail. For more information, see Scheepers (2014).

3.3 Investigation of concordance lines and identification of errors

The investigation of the corpora was driven by two research questions:

(i) How does students' production (in terms of frequency) of selected MWUs compare with the production of these MWUs by expert writers, within and across courses?

(ii) How does Literature students' production (in terms of frequency and deviance) of selected MWUs compare with the production of these MWUs by Law students?

Multiword units containing the verbs *make*, *take* or *have* were extracted from the corpora using Wordsmith Tools (WST) (Version 5) (Scott 2008). The investigation of concordance lines and MWUs was done manually. The study was essentially corpus-driven, although it did include an element of corpus-based methodology, in that the researcher had made the prior decision to investigate 'interesting words', namely the high-frequency verbs *have, make* and *take* (Biber 2009:276). The corpora were not annotated, and the analysis of the concordance lines was inductive rather than deductive: the focus was on the patterns and clusters featuring these verbs that emerged from a manual investigation of concordance lines.

In addressing these research questions, Wang and Shaw's (2008) steps were followed. Only the steps pertaining directly to this article are explained here.

The first step was to generate wordlists for the verbs in question, using the *WordList* application of WST. Although individual word frequency was of less interest than that of multiword items, it was important to establish at the outset whether the three verbs were in fact frequent in these corpora.

In the second step, WST's *Concord* application was used to generate a separate concordance for each word form of each selected verb; that is, the "inflectional morphemes" (Biber 2006:34) of the base word or lemma, e.g. *have*, *has*, *had*.

Once the concordance lines had been generated, the functions and categories of use of these verbs were manually classified, using Biber et al. (1999) as a guide, in order to isolate the focus of the study – the delexical uses of the verbs. Based on Algeo (1995) and Nesselhauf (2003, 2005), combinations were considered to be core delexical MWUs if they featured *have*, *make* or *take* occurring

- with an eventive noun, where the verb carried little lexical weight or was semantically empty;

- where there was a verb, identical[4] in form, that could replace the whole combination, e.g. *let's have a look* could be replaced with the verb *look*; and

[4] "Identical" is to be interpreted flexibly to accommodate necessary grammatical changes driven by tense and person.

- where the eventive noun was preceded by an indefinite article (a/an). For example,
  study the literature if you do not ***have a love for*** *reading per se*
  purpose as a whole is to ***make a comment*** *on the problem* (Exp Lit).[5]

Combinations were considered *pseudo* delexical MWUs when they failed in some way to fulfil the requirements for core delexical MWUs:

- where the replacement verb was not identical to the noun, but morphologically related, as in *he made a decision* and *he decided* (Langer 2004:17), an example of what Algeo (1995) terms "affixation", but which is usually referred to as "derivation". Adding an affix to a word changes the word's part of speech: thus, the verb *decide* can be changed to a noun, *decision,* by the addition of a suffix *–ion.* Examples of this type from the Student corpus included *He* ***makes a very interesting observation*** (Stud. Lit);

- where there was "a flaw in correspondence between the expanded predicate and a corresponding simple verb" (Algeo 1995:206) caused by: modulation (change of prosodic phonemes) and phonological modification (change of segmental phonemes), e.g. *make a prótest = protést, take a breath = breathe* (Algeo 1995:205); pluralisation; the use of the definite article instead of the indefinite article; omission of the article; no corresponding single-word verb in everyday use, e.g. *have a game*, *make an effort, have an affair*; and there being only an equivalent non-cognate single-word verb, e.g. *take cover* = *hide* (Algeo 1995:206);

- where the eventive noun was morphologically related to a simple verb, but the delexical MWU differed semantically from that verb, e.g. *make love ≠ to love*, *have a bite ≠ to bite* (Algeo 1995:206);

- where the corresponding simple verb was passive rather than active, e.g. *have a fright* = *be frightened* (ibid.).

Once these MWUs had been identified, they were counted and the numbers in each corpus were compared. Rayson's log-likelihood calculator was used to identify significant differences between corpora.[6]

In the next step, deviant MWUs and the errors they contained were categorised and quantified. A distinction was made in this study between 'deviation' and 'error'. 'Deviation' refers to those MWUs that were in some way problematic, while 'error' refers to the specific way in which such MWUs were deviant. Of the delexical MWUs in the Student corpora, varying proportions were deviant. This study investigated all types of errors in collocations, whether grammatical or lexical, but only those that were integral to the MWU itself. In other words, it considered only the verb, article and noun elements of the collocation and the particles immediately preceding or following the noun.

[5] The source applies to the concordance line at the end of which it occurs and to any following lines. Where a different source is used in the same set of examples, a space is inserted and the rule applied accordingly. Corpora are referred to as "Exp" (Expert) or "Stud" (Student) "Lit" (i.e. Literature) or "Law".

[6] Rayson's log-likelihood calculator: http://ucrel.lancs.ac.uk/llwizard.html

The classification of errors was based on Nesselhauf's "Types of mistakes in collocation" (2003:232), but with some sub-categorisations and the additional category of adjective (ADJ) added. Seven main categories of error were identified:

(i) Adjective (ADJ)

Examples included (1) below, where the adjective is not a fully appropriate collocate with *living*:

(1) dealers are **making a wealthy living** out of (Stud Law)

(ii) Determiner (D)

Deviations in the determiner included errors in the article, as in example 2, where the indefinite article *a* is missing after *has* (*has a life*); (3) incorrect article, where the wider context indicated that *a* should be replaced by the definite article *the* (*has the connotation*); or (4) present but inappropriate article. In example 4, besides the concord error in the verb, the article is present but inappropriate in *have the disgrace*:

(2) They make the city of London **has life** like a living being. (Stud Lit)

(3) The Madonna of Excelsior **has a connotation** of the women

(4) Melanie's father. Lurie says he **have the disgrace** for his who

Deviations in the determiner included errors to do with the pronoun. In example 5, the pronoun, in this case the possessive pronoun *her,* is missing:

(5) alerting Agnes to stand and **make voice** be heard. (Stud Lit)

In example 6, a pronoun is present but incorrect, as *his* should be replaced by *her*:

(6) a councillor meanwhile *she* is **making *his* way** to the top (Stud Law)

Finally, there were errors to do with the demonstrative, as in example 7 below, where *this* should have been replaced by *these* to agree in number with the plural *notes*:

(7) they come with this technology of **making *this* notes** this lead to (Stud Law)

(iii) Noun (N)

Errors concerned number, where the plural was used where the singular was required, and vice versa, as in example 8, where *mistake* should be plural to agree with *a lot of*:

(8) got in this world. David **made a lot of *mistake*** in life (Stud Lit)

(iv) Preposition (P)

Errors included prepositions present though incorrect, as in example 9 below, where *off* should be replaced with *of*; and in example 10, where *to* should be replaced with *for*. These prepositions were determined by the MWU in each case and so were identified as errors:

(9) that she and her family **would be taken care *off*** if she does do (Stud Lit)

(10) He has changed because he **has sympathy *to*** his daughter

(v) Structure (S)

Errors of syntax occurred, as in example 11, where the structure of the whole expression was incorrect. In this case, interpretation was not straightforward. The italicised expression could be replaced with *If a bribe is what it takes?* or *If it takes a bribe (to achieve something…)*:

(11) anything achieve that. *If it* ***take to bribe*** *somebody* in order (Stud Law)

(vi) Stretched verb construction (SVC)

Deviations comprised examples such as (12), where a simple verb (*trust*) would have been more appropriate than an SVC:

(12) this fellow Uriah. He **does not have any trust** towards this (Stud Lit)

(vii) Verb (V)

Deviations included: tense, in most cases the overuse of the progressive aspect present tense; concord; errors in collocation; and errors in verb choice. Example 13 is an error in the tense of the verb, where the progressive aspect present tense should be replaced by the simple present tense *has*:

(13) lines sestet where the poet *is* **having a solution** *of* his problem (Stud Lit)

Example 14, in addition to other errors, contains an error of concord (subject-verb agreement), where *have* should be replaced by *has*:

(14) and the people was dying. He **have no hope** *of* the universe (Stud Lit)

In example 15, the student has collocated *has* incorrectly with *contribution*. *Contribution* collocates with *make*:

(15) today's newspaper. Immigration **has had a great contribution** (Stud Law)

In example 16, the choice of verb is incorrect (*heed* should have been used rather than *hid*).

(16) away, advice that he **does not take hid** of. Gatsby could also (Stud Lit)

In several cases, more than one error occurred in an MWU, sometimes with the cumulative effect of making interpretation difficult. In all cases, each error was coded and explained. In cases where one error caused another, as in example 17:

(17) Yes I agree some of them **have connection** with our guys (Stud Law),

where the indefinite article after *have* is missing and there is an error in the noun (*connection* should be plural), only one error was counted; in this case the absence of a determiner. Decisions in such cases were not straightforward, but here the rationale had to do with the key role that the determiner plays in the definition of the focal structure, the delexical MWU.

4. Results and discussion

The sections below address the research questions in turn, and results are discussed accordingly.

4.1 Frequency of MWUs in Student and Expert corpora (Q 1)

The frequency of MWUs in the Student corpora was compared to their frequency in the Expert corpora. Subsequently, the frequencies of MWUs in the two Student corpora were compared. Table 2 below presents these frequencies and the number of occurrences of each verb in the corpora:

Table 2: Delexical MWUs – all corpora

Verb (V)	**Corpus**	**Core delexical MWUs (CD)**	**Pseudo delexical MWUs (PD)**	**Total**	**Total no. of occurrences of V in corpus**	**% of V made up by delexical MWUs**
HAVE	Expert Lit	20	124	144	1345	10.70
	Student Lit	24	207	231	1626	14.20
	Expert Law	3	40	43	399	10.77
	Student Law	3	74	77	790	9.74
MAKE	Expert Lit	15	136	151	350	43.14
	Student Lit	5	117	122	493	24.74
	Expert Law	2	61	63	135	46.66
	Student Law	1	43	44	158	27.84
TAKE	Expert Lit	6	63	69	222	31.08
	Student Lit	13	156	169	341	49.56
	Expert Law	6	25	31	65	47.69
	Student Law	4	55	59	140	42.14

These results can be viewed from two perspectives. The first is a narrow perspective: that is, as far as the number of MWUs according to the occurrence of the three verbs in the corpora is concerned, the verbs *make* and *take* proved to be the most productive of the delexical combinations in this study, with higher proportions of MWUs (i.e. relative overuse) with *take* in the Student corpora than in the Expert corpora. This is reflected in the last column of the table. From a broader perspective, however – that is, using the log-likelihood (LL) calculation based on the occurrence of MWUs with reference to whole corpora – significant differences between the totals for all delexical MWUs (column 5 of the table) occurred between the Student Lit and the Expert Lit corpora in the case of *have* (LL = 21.34, p < 0.0001) and *take* (LL = 44.46, p < 0.0001), with very significantly more occurrences in the Student corpus in both cases.[7] This overuse of *take* as a delexical verb was partly the result of the excerpts used in the Literature examination paper, which contained expressions such as *take advantage (of)*, which made up almost half (48.9%) of all MWUs in the Student Lit corpus. In the Law corpora, the Student corpus featured very significant underuse of the delexical *make* relative to the Expert corpus (LL = 10.93, p < 0.0001).

There were no significant differences in the delexical use of the three verbs between the two Expert corpora. In the case of the Student corpora, however, Literature students produced significantly more delexical MWUs than their Law peers, and there was significant overuse of *have* as a delexical verb (both core and pseudo) in the Lit corpus relative to the Law corpus (LL = 5.07, p < 0.05).

4.2 Deviant MWUs and errors (Q2)

This section addresses the differences between the two Student corpora in terms of numbers of deviant MWUs and numbers and types of errors. As the focus was on student errors or deviations, the Expert corpora are not discussed further.

At this point, because the number of deviant core delexical MWUs was so small (only six in total), the categories of core and pseudo were combined to form one category, delexical MWUs.

Table 3: Student corpora – numbers of deviant delexical MWUs and errors

Verb	**Corpus**	**CD**	**PD**	**Total delexical MWUs**	**Deviant MWUs**	**% Dev MWUs**	**Errors**
HAVE	Stud Lit	24	205	229	33	14.4	51
	Stud Law	3	72	75	15	20.0	16
MAKE	Stud Lit	5	117	122	12	9.8	13
	Stud Law	1	41	42	17	40.4	22
TAKE	Stud Lit	13	156	169	12	7.6	13
	Stud Law	4	55	59	20	45.7	27

Key: CD: core delexicals; PD: Pseudo delexicals

The percentage of deviant MWUs for each verb varied from as low as 9.8% for *make* in the Student Lit corpus to as high as 45.7% for *take* in the Student Law corpus. The Lit corpus

[7] The higher the LL value (G2), the more significant the difference between two frequency scores. For this study, a G2 of 3.8 or higher was significant at the level of p < 0.05 and a G2 of 6.6 or higher was significant at p < 0.01.

- 95th percentile; 5% level; p < 0.05; critical value = 3.84
- 99th percentile; 1% level; at p < 0.01; critical value = 6.63
- 99.9th percentile; 0.1% level; p < 0.001; critical value = 10.83

contained a lower percentage of deviant MWUs in general (LL = 14.44, $p < 0.001$) and very significantly fewer errors (LL = 16.39, $p < 0.0001$) relative to the Law corpus. This difference between corpora was reflected in the results for both *make* (deviant MWUs: LL = 10.84, $p < 0.001$; errors: LL = 16.64, $p < 0.0001$) and *take* (deviant MWUs: LL = 9.53, $p < 0.01$; errors: LL = 22.71, $p < 0.0001$), with the *Lit* corpus producing significantly fewer deviant MWUs and errors relative to the Law corpus. Differences between corpora for *have* were not significant. As far as the three verbs were concerned, *take* appeared to cause particular difficulty for Law students, with 45.7% of MWUs with this verb being deviant in some way. Law students showed a lack of awareness of the collocational restrictions on both *take* and *make*, illustrated in the examples below:

(18) former president of Justice **was making a serious corruption** (Stud Law)

(19) the rich because poverty **takes part**. In most cases of

(20) of some Africans also **were taken rescued** by the Spanish

These findings underline what other studies have found – that combinations featuring high-frequency verbs used delexically are notoriously difficult for learners to master (Altenberg and Granger 2001; De Cock and Granger 2004; Kaszubski 2000; Lee and Chen 2009; Wang and Shaw 2008).

In order to explain these differences in the two Student corpora, one must consider the types of errors made. Table 4 below provides a breakdown of errors per category.

Table 4: Types of error

Verb	**Corpus**	**ADJ**	**D**	**N**	**P**	**S**	**SVC**	**V**	**Total errors**	**Total deviant delexical MWUs**
HAVE	Stud Lit	1	10	-	10	-	8	22	51	33
	Stud Law	-	4	1	4	-	-	7	16	15
MAKE	Stud Lit	-	4	2	1	-	-	6	13	12
	Stud Law	1	7	1	-	-	1	12	22	17
TAKE	Stud Lit	1	1	2	3	-	1	5	13	12
	Stud Law	1	6	2	2	1	2	13	27	20
Total deviant %	Stud Lit	2 2.5%	15 19.4%	4 5.1%	14 18.1%	-	9 11.6%	33 42.8%	77	57
	Stud Law	2 3.0%	17 26.1%	4 6.1%	6 9.2%	1 1.5%	3 4.6%	32 49.2%	65	52

There were relatively few errors in the adjective category (two in each corpus). In example 21 below, the error lies in the fact that the student has used the adverb *totally* in place of the correct adjective *total,* although it could be argued that although the target adjective (*total*) does fall within the MWU, the actual word used (*totally*) does not:

(21) Babamukuru's approval, she totally **had her dependence**. Even when (Stud Lit)

The error in example 22, explained as an error of adjectival collocation, is more complex in that *living* would collocate more correctly with *good* or *successful* in this context:

(22) with drugs. Drug dealers **are making a wealthy living** out of (Stud Law)

Both these errors suggest a lack of awareness of both the grammatical and the phraseological conventions of the language. The error in the adjective is particularly interesting, as it indicates that students may have difficulty with more than simply verb-noun collocations; although confirmation of this would require a further search for adjective combinations in the corpora.

Errors in the determiner occurred in the use of the central determiners (articles, demonstratives, and possessive determiners) (Biber et al. 1999:258). The aspect that caused the most trouble for students was the article, with the majority of determiner errors falling into this category: 86% in the Lit corpus and 78% in the Law corpus. Examples included cases where the article was omitted, such as (23):

(23) the rich people because poverty **takes part**. In most cases of (Stud Law);

or where a definite article was used instead of the indefinite article, as in (24):

(24) The neighbours **are having the differences** between them (Stud Law)

Errors in article use are particularly common in the variety of English spoken by the majority (56.7%) of students in the study, Black South African English (BSAE): 19.4% of errors in the Student Lit corpus (comprising 46.7% black students) and 26.1% in the Student Law corpus (65% Black students) fell into this category. Van Rooy (2013:12) confirms that the errors listed here are three possible ways in which article use in this variety differs from native speaker varieties: articles are left out altogether; articles are inserted where they would not be used at all in NS English; or articles are muddled, that is, substituted for each other (Greenbaum and Mbali 2002:241-3). In De Klerk's (2006a:146) analysis of a corpus of Xhosa English, she found evidence of a "loss of distinction between mass and count nouns", with attendant use of both the definite and indefinite article with non-count nouns, as in example 25:

(25) rich people some of them **are taking an advantage** of poor.

De Klerk (2006a) also found examples of the omission of articles and the insertion of inappropriate articles, as in this study. Minow (2010), focusing on the omission of definite and indefinite articles, showed that the insertion of an article where native varieties would not use one was the most frequent difference in BSAE.

Other examples in the category of determiner, though less common, involved errors in the use of the possessive pronoun, as in (26):

(26) at all but let justice and law **takes its effects** (Stud Law),

where the plural form of the possessive pronoun *its*, i.e. *their*, should have been used.

The difference in the number of determiner errors in the two corpora was significant (LL = 6.84, p < 0.01), with the Law students making significantly more errors in this category than the Literature students. Most errors in the determiner occurred in the use of *have* in the Lit corpus (66.6%) and in the use of *make* in the Law corpus (41.1%). Law students made significantly more errors than Lit students in the use of *make* (LL = 5.01, p < 0.05) and very significantly more for *take* (LL = 9.12, p < 0.01) in this category.

While Nesselhauf (2005:71) found that the noun formed the second-most frequently deviant element in her corpus, and in many cases the whole collocation was "inappropriate" (including SVCs, which should have been single-word verbs), in this study, nouns caused relatively fewer problems for student writers (5.1% of errors in the Student Lit corpus involved the noun, and only a slightly higher 6.1% in the Student Law corpus). All errors in this category concerned number, either in verb-noun agreement or article agreement, as in example 27:

(27) the beauty of the City. This **makes a contrasts** with its use (Stud Lit),

or where uncountable nouns were used incorrectly in the plural, as in example 28:

(28) practices. It is the rich who **are having these accesses** in most (Stud Law)

Differences between the two corpora in the number of errors in this category were not significant in the case of any of the verbs or overall.

In contrast to the noun, the preposition caused students considerable difficulties, particularly in the case of bound prepositions. Errors in prepositions made up the third-largest group, after verbs and determiners. There were no significant differences between corpora in the number of preposition errors made with each of the three verbs. Examples of errors in preposition use included:

(29) greedy. Greed is from the devil: **take a look** *of* the following (Stud Law)

(30) Lack of education therefore **has a direct correlation** *to* the (Stud Law)

In both these examples, the error occurs because the preposition is bound by the context, providing further evidence of a lack of awareness of the idiomatic nature of certain combinations in English. In fact, 12 of the 20 preposition errors (60%) occurred in cases where the preposition was bound.

The category of structure is defined as "syntactic structure wrong" (Nesselhauf 2003: 232). Only one error fell into this category:

(31) anything achieve that. If it **take to bribe** somebody in order (Stud Law)

As explained in Section 3, this error makes the entire expression difficult to explain in terms of any of the other categories and as such requires a more general category expressing the overall ambiguity of the combination.

Errors arising from the use of a delexical combination rather than a single-word verb fell into the SVC category. Most errors occurred in the Lit corpus in the use of *have* (66.6%), but this category made up only 8.4% of the total number of errors in the two corpora. Examples included:

(32) definitely be why they mostly all **have a dependency** on him to work (Stud Lit)

Although this expression with an article appears 21 times in the British National Corpus (BNC), in the Lit corpus a single-word verb such as *depend* would have been more appropriate. In addition, *dependency* is uncountable and does not take an article. In example 33 below, the single verb *trust* would have been more appropriate than an MWU:

(33) this fellow Uriah. He does not **have any trust** *towards* this man (Stud Lit)

Some constructions featured an underuse of academic words, as in example 34 below:

(34) rate due to the **movement they make** to other places because (Stud Law)

Here the context indicates that the noun *migration*, or even *immigration*, would have been more appropriate. The effect of such errors is to make students' writing sound laboured and non-nativelike, and suggests a limited vocabulary.

The largest group of errors in both corpora fell into the verb category: 42.8% of all errors in the Lit corpus and 42.9% in the Law corpus. Differences between the two corpora were significant for *make* (LL = 9.76, $p < 0.01$) and highly significant for *take* (LL = 13.02, $p < 0.001$). The difference in the total number of verb errors in the two corpora was also significant (LL = -9.57, $p < 0.01$), with more errors in the Law corpus. These findings reflect those of Nesselhauf (2003, 2005). She explains her findings with reference to the commonly held belief that verbs are the most difficult words for learners to master. Nesselhauf (2005:77) found that her learners confused "high-frequency Germanic verbs, such as *take* and *make*, *get* and *give* etc.", adding weight to this study's premise that these words can be difficult for learners to master.

These findings support those of other studies (Altenberg and Granger 2001; Kaszubski 2000; Laufer 1991; Lee and Chen 2009; Wang and Shaw 2008; Yan 2006). Altenberg and Granger (2001:189), for instance, found that even at an advanced proficiency level, learners had great difficulty with these verbs, and when the verb was used delexically this difficulty was compounded.

In this study, the majority of all verb errors occurred in the collocation category (38.4%), highlighting students' limited awareness of the collocational restrictions governing these verbs:

(35) business suits who choose to **take part** in corruption, who (Stud Law)
[*corruption* collocates with *commit*, not *take part in*]

(36) good and in deep humility. He **takes conversation** with other (Stud Lit)
[*conversation* collocates with *make*, not *take*]

(37) with different eyes thus he **makes a conclusion** that will
[*conclusion* collocates with *reaches*, *arrives at*, not *makes*].

The Law corpus produced significantly more errors than the Lit corpus in this category (LL = 9.04, $p < 0.01$). This was an aspect of verb use that Law students found particularly difficult, especially in the case of collocations with nouns such as *corruption* (9 errors) and *contribution* (2 errors), which made up 73.3% of all collocation errors in the Law corpus.

There was only one example of an incorrect choice of verb, the use of *hid* in place of *heed* in the example below. This may in fact have been a spelling error; there may be no distinction between long and short vowels such as *heed* and *hid* in the pronunciation of many BSAE speakers:

(38) away, advice that he does not **take *hid*** of. Gatsby could also (Stud Lit)

Errors of verb tense and concord made up the bulk of the remaining verb errors in the two corpora, with no significant differences between corpora. ‘Tense’ is used here in the more traditional sense of the term to include aspect. However, to be more specific, although tense and aspect both “relate primarily to time distinctions in the verb phrase” (Biber et al. 1999:460), tense describes the time at which an action takes place, either in the past or in the present, while aspect denotes whether the activity or state is ongoing or completed. In this study, tense errors in the verb predominantly involved examples where the progressive aspect was used in a non-standard manner. In Standard English (SE), the present progressive aspect is used to describe actions that are currently in progress or that are about to take place in the near future (Minow 2010:129). However, “[t]he progressive aspect has a long history of scholarly attention in BSAE and many other New Englishes” (Van Rooy 2013:11). Although stative verbs such as *have* that refer to “unchanging conditions” are not usually used in the progressive aspect (Richards and Schmidt 2002:34, 513), corpus analysis (De Klerk 2006a; Van Rooy 2006, 2013) has confirmed the “extension of the progressive to stative verbs” in some learner varieties. Although Minow (2010:144) found that in her Xhosa data “the frequency of the progressive decreases with increasing proficiency”, which suggests that the extension of the progressive may be “a learner phenomenon” which will disappear as proficiency increases, Van Rooy (2013:11) observes that his data (Van Rooy 2006) revealed that the “underlying semantics of the construction is consistently different from the native speaker prototype of a dynamic event with a limited duration”. Most uses in his data reflected extended duration; in such cases, the “construction is equally compatible with dynamic and stative predicates” (2013:11). In BSAE, unlike in NS and foreign language uses, the “temporariness, imminent change and the activity being ongoing or foregrounded at some temporal reference point are not central to the meaning” (Van Rooy 2011:196). Van Rooy’s examples from the BSAE corpus “form a coherent linguistic construction”, differing essentially from SE (ibid.). As noted above, it is likely that a significant number of students in both groups, as mother-tongue speakers of an indigenous language, would have revealed features of BSAE in their writing.

Although the examples categorised as errors in this corpus of student writing could thus be regarded as aspects of an increasingly acceptable language variety, I indicated them as problematic because in the particular context the construction could be regarded as non-standard and not what is required in academic writing. The simple present or simple past tense would have been preferable in the examples below:

(39) neighbours. The neighbours are **having the differences** between (Stud Lit)

(40) lines sestet where the poet **is having a solution** *of* his problem or hill, were untouched.

(41) It is the rich who **are having these accesses** in most (Stud Law)

(42) former president of Justice was **making a serious corruption**,

As far as errors of concord are concerned, as in the case of errors in noun and article use, these highlight students' lack of awareness of agreement in general and suggest a lack of depth of vocabulary knowledge. The Law corpus featured more errors relative to the Lit corpus, but this was not significant. De Klerk (2006a:43) observes that a "tendency to simplify concord […] is a frequently remarked-on feature of BSAE" (De Klerk 2006b) and these findings certainly attest to this. Subject-verb agreement errors, a typical feature of the English spoken by Afrikaans mother-tongue speakers (Coetzee 2009), were also frequent. There were 25 Afrikaans-speaking students in the Lit corpus and 26 in the Law group. But learners also had difficulties with the use of determiners, particularly articles and pronouns. These too are features of English which are known to cause difficulties for speakers of indigenous South African languages, where the pronoun does not always exist as an independent word and where there may be no article. These findings support the work of other researchers of South African varieties of English (Coetzee 2009; De Klerk 2006a, 2006b; Van Rooy 2006, 2013) and learner errors (Nel and Muller 2010; Nel and Swanepoel 2010), suggesting that differences between the two corpora may partly have resulted from the influence of learners' L1.

In all cases except preposition and SVC errors, Student Law writers made more errors than the Literature students: the differences are significant in the determiner and verb categories, and while both these aspects are known to cause learners difficulties, it is clear that Law students found verb use particularly difficult and showed less awareness of collocational restrictions than the Literature students. Law students produced significantly more deviant MWUs (LL = 13.56, $p < 0.001$) as well as significantly more errors (LL = 13.94, $p < 0.001$).

The errors in the verb bear out what has been observed in the marking of assignments and examination scripts, where the overuse of the progressive aspect and concord deviations are common problems in verb use among these students. To sum up, errors in both Student corpora reflected a lack of collocational awareness and restriction and the sometimes arbitrary nature of this restriction. Errors reflected a limited awareness of the rules of usage of high-frequency verbs as well as a paucity of lower-frequency and academic words, that is, a lack of depth of vocabulary knowledge. An example such as *they mostly all have a dependency on him to work*, for instance, reflects limited awareness of the way in which, as a word's function changes, so in many cases does its spelling and form. This lack of awareness of inflections and derivations is a characteristic of much of the writing by learners such as these.

This is not unique to these learners, however. Yan (2006:40-41) found that her Chinese students "are always allowing delexical *do* [my emphasis] more freedom to collocate with a wide range of nouns" and that "learners do not only overuse delexical structures, they also misuse them". In the study discussed in this article, the data revealed an overuse of unnativelike collocations such as *make + corruption*, *have + solution*, *have + contribution* and *make + conclusion*. Furthermore, like Farghal and Obiedat (1995:321), who found that their Arabic subjects' "unawareness of colloquial restrictions of lexical items" led them to produce deviant collocations, this 'unawareness' was clear in many of the errors in this study.

In this study, students produced, or attempted to produce, more delexical MWUs than the Expert writers. Thus, in contrast to Altenberg and Granger's (2001) findings, they tended to *overuse* delexical combinations, particularly where single-word verbs would have been preferable. However, they also misused these combinations. A lack of awareness of collocational restrictions was, like that of Farghal and Obiedat's (1995) students, particularly evident from the MWUs produced in the Law corpus. Like learners in the majority of studies discussed here (Kaszubski 2000; Altenberg and Granger 2001; Gilquin 2007; Laufer and Waldman, 2011; Yan 2006), though in contrast to Nesselhauf's and Howarth's studies, these students did have difficulty with delexical combinations.

As far as the types of errors were concerned, numerically, *take* and *make* were most productive. The differences between the two Student corpora were more complex, however. Both the proportion of delexical MWUs and the proportion of deviant MWUs and errors reflected marked differences between writers from the two courses. Altogether, the Student Lit corpus produced significantly more delexical MWUs than the Student Law corpus, but significantly fewer of these were deviant and they made significantly fewer errors. The Lit students wrote extended texts from the beginning of their course and were required to read several full-length texts – novels, plays and poetry. The focus of the course is on *how* students write, not simply on *what* they write. Law students, in comparison, wrote very little in the way of extended texts during the semester, and their prescribed reading comprised mostly legal cases. Errors in the Student Law corpus particularly reflected what Howarth (1998b:186) describes as a lack of awareness of the "existence of the central area of the phraseological spectrum between free combinations and idioms", that is, restricted collocations. Errors stemming from such an unawareness were reflected in this study in the many errors in the collocation of *corruption* in this corpus, for instance.

Despite these differences between corpora, however, all students revealed gaps in their collocational awareness and made other errors in the MWUs they produced. This may partly be the result of a lack of practice in writing and limited extended reading in general; although Literature students engaged in more extended reading and writing in their English module than Law students, relatively speaking they were required to do very little writing during the semester. The fact that there is little teaching of grammar and language in these courses may also have exacerbated this situation. A further element of the deviations in the Student corpora is that many errors appear to have become habitual among learners. The errors made by both groups of writers illustrate features that are common in student writing at this university: article and tense markers are frequently omitted, the progressive aspect is commonly overused or used inappropriately, and pronouns are often used interchangeably. The fact that there is not an established culture of reading in South Africa only exacerbates this situation.

5. Concluding remarks

Analysis of the corpora in this study revealed that, in contrast to studies such as Altenberg and Granger's (2001), students tended to *overuse* delexical combinations, particularly where single-word verbs would have been preferable, and produced more delexical MWUs than the Expert writers. But they also misused these combinations, particularly in the case of the Law students, suggesting a lack of awareness of collocational restrictions. Thus, like learners in most of the studies discussed in the above sections, these students did have difficulties with delexical combinations.

The Student Lit corpus produced significantly more delexical MWUs than the Student Law corpus, but significantly fewer deviations and errors. Nevertheless, all students revealed gaps in their collocational awareness and made other errors in the MWUs they produced. The verbs *take* and *make* were most prone to error, and in both student corpora the majority of errors were in the verb element of the MWU, with the Law students producing significantly more errors than the Literature students. The largest group of errors in the verb occurred in the collocation category, highlighting students' limited grasp of the collocational restrictions governing these verbs.

It became clear from the types of errors in the MWUs produced by students that many had not developed a depth of knowledge of high-frequency words. Errors of collocation in particular reveal the importance of a deeper knowledge of high-frequency verbs; the errors suggest a lack of awareness of how such seemingly simple words behave together with others, and the restrictions that are frequently imposed by their collocational properties. Thus, as in a study by Lee and Chen (2009), learners appeared to be unaware of the subtleties involved in the use of these words. This lack of depth of knowledge of high-frequency words was combined with a lack of knowledge of lower frequency words. In other words, there were deficits in both breadth and depth of vocabulary knowledge. A lack of awareness of inflections and derivations also made students' writing sound unnativelike. As Laufer and Waldman (2011:666) observe, the fact that "use of incorrect collocations makes people sound odd but does not impair communication altogether" means that "language accuracy" may be neglected to the detriment of the development of "collocational knowledge".

In recent years, an increasing amount of research has investigated these high-frequency words (e.g. Altenberg and Granger 2001; Gilquin 2007; Kaszubski 2000; Lee and Chen 2009; Liu and Shaw 2001; Nesselhauf 2004, 2005; Wang and Shaw 2008; Yan 2006). The findings of this study add support to evidence that these words may indeed be the "*bête noire*" of learners (De Cock and Granger 2004:233).

There is certainly scope for further research in this domain in the South African context. Closer examination of the relationship between the use of MWUs, especially those containing high-frequency words, and reading comprehension, vocabulary levels and academic proficiency could provide more insight into the difficulties students such as those in this study have with vocabulary in general and with MWUs in particular. There is also a need for more research into vocabulary knowledge in African languages and the use of high-frequency words in African language corpora. In addition, the perspective on depth of vocabulary knowledge in this study was a narrow one. More work could be done in this area; for instance, by comparing MWUs to other measures of depth, such as the Word Associates Test (WAT) (Read, 1993, cited in Akbarian 2010).

These findings have implications for students at university: if they wish to compete in the academic milieu, they must be able to write in a way that is accurate and stylistically appropriate. As Hyland (2013:54) observes, "English has emerged as the international language of research and scholarship". Although these high-frequency words are often regarded as less important than academic words, and although errors in their use may not affect communication, such errors can have a cumulative effect on students' production, affecting the quality of their writing (Lee and Chen 2009:121). A greater focus on these 'little' high-frequency words and their collocations in our teaching is thus crucial.

References

Akbarian, I. 2010. The relationship between vocabulary size and depth for ESP/EAP learners. *System,* 38: 391-401. doi:10.1016/j.system.2010.06.013

Algeo, J. 1995. Having a look at the expanded predicate. In B. Aarts and C.F. Meyer (eds.), *The verb in contemporary English*: *Theory and description*. Cambridge: Cambridge University Press. pp. 203-217.

Altenberg, B. and S. Granger. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics* 22(2): 173-195. doi:10.1093/applin/22.2.173

Biber, D. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam/Philadelphia: John Benjamins.

Biber, D. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3): 275-311. doi:10.1075/ijcl.14.3.08bib

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan. 1999. *The Longman grammar of spoken and written English*. London: Longman.

Coetzee, W. 2009. *Language errors in the use of English by two different groups of Afrikaans first language-speakers employed by Nedbank: An analysis and possible remedy.* Unpublished MA thesis. Stellenbosch: University of Stellenbosch.

Coxhead, A. 2000. A new academic wordlist. *TESOL Quarterly* 34(2): 213-238. doi:10.2307/3587951

De Cock, S. and S. Granger. 2004. High frequency words: The bête noire of lexicographers and learners alike. In G. Williams and S. Vessier (eds.), *Proceedings of the Eleventh Euralex International Congress*. Université de Bretagne-Sud: Lorient. pp. 233-243.

De Klerk, V. 2006a. *Corpus linguistics and world Englishes: An analysis of Xhosa English*. London, New York: Continuum.

De Klerk, V. 2006b. The features of "teacher talk" in a corpus-based study of Xhosa English. *Language Matters* 37(2): 125-140. doi:10.1080/10228190608566257

Erman, B. and B. Warren. 2000. The idiom principle and the open choice principle. *Text* 20(1): 29-62. doi:10.1515/text.1.2000.20.1.29

Farghal, M. and H. Obiedat. 1995. Collocations: A neglected variable in EFL. *International Review of Applied Linguistics in Language Teaching* 33: 315-331. doi:10.1515/iral.1995.33.4.315

Gilquin, G. 2007. "To err is not all." What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift für Anglistik und Amerikanistik* 55(3): 273-291. doi:10.1515/zaa.2007.55.3.273

Granger, S. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A.P. Cowie (ed.), *Phraseology: Theory, analysis and applications.* Oxford: Clarendon Press. pp. 145-160.

Greenbaum, L. and C. Mbali. 2002. An analysis of language problems identified in writing by low achieving first-year students, with suggestions for remediation. *South African Linguistics and Applied Language Studies* 20: 233-244. doi:10.2989/16073610209486313

Hasselgren, A. 1994. Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 2: 237-260. doi:10.1111/j.1473-4192.1994.tb00065.x

Howarth, P. 1998a. Phraseology and second language proficiency. *Applied Linguistics* 19(1): 24-44. doi:10.1093/applin/19.1.24

Howarth, P. 1998b. The phraseology of learners' academic writing. In A.P. Cowie (ed.), *Phraseology: Theory, analysis and applications.* Oxford: Clarendon Press. pp. 161-186.

Hyland, K. 2013. Writing in the university: Education, knowledge and reputation. *Language Teaching* 46(1): 53-70. doi:10.1017/s0261444811000036

Kaszubski, P. 2000. *Selected aspects of lexicon, phraseology and style in the writing of Polish advanced learners of English: A contrastive, corpus-based approach*. Doctoral dissertation. Poznań: Adam Mickiewicz University.

Langer, S. 2004. A formal specification of support verb constructions. In S. Langer and D. Schnorbusch. (eds.), *Semantik im Lexikon*. Tübingen: Narr. pp. 179-202. Available online: http://129.187.148.72/download/publikationen/05stefan_langer_dgfs.pdf (Accessed 25 May 2012).

Laufer, B. 1991. The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal* 75(4): 440-448. doi:10.1111/j.1540-4781.1991.tb05380.x

Laufer, B. and I.S.P. Nation. 1999. A vocabulary-size test of controlled productive ability. *Language Testing* 16(1): 33-51. doi:10.1177/026553229901600103

Laufer, B. and T. Waldman. 2011. Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning* 61(2): 647-672. doi:10.1111/j.1467-9922.2010.00621.x

Lee, D.Y.W. and S.X. Chen. 2009. Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing* 18: 149-165. doi:10.1016/j.jslw.2009.05.004

Léon, J. 2007. Empiricism versus rationalism revisited: Current corpus linguistics and Chomsky's arguments against corpus, statistics and probabilities in the 1950-1960s. In S. Matteos and P. Schmitter (eds.) *Linguistische und epistemologische Konzepte – Diachron.* Munster: Nodus Publikationen. pp. 157-176. Available online: http://htl.linguist.univ-paris-diderot.fr/leon/empiricism2007.pdf (Accessed 2 February 2011).

Liu, E.T.K. and P.M. Shaw. 2001. Investigating learner vocabulary: A possible approach to looking at EFL/ESL learners' qualitative knowledge of the word. *IRAL, International Review of Applied Linguistics in Language Teaching* 39(3): 171-194. doi:10.1515/iral.2001.001

Live, A.H. 1973. The take-have phrasal in English. *Linguistics* 95: 31-50. doi:10.1515/ling.1973.11.95.31

Martinez, R. and N. Schmitt. 2012. A phrasal expressions list. *Applied Linguistics* 33(3): 299-320. doi:10.1093/applin/ams010

Minow, V. 2010. *Variation in the grammar of Black South African English*. Frankfurt Am Main: Peter Lang.

Nel, N. and H. Muller. 2010. The impact of teachers' limited English proficiency on English second language learners in South African schools. *South African Journal of Education* 30(4): 635-650.

Nel, N. and E. Swanepoel. 2010. Do the language errors of ESL teachers affect their learners? *Per Linguam* 26(1): 47-60. doi:10.5785/26-1-13

Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24(2): 223-242. doi:10.1093/applin/24.2.223

Nesselhauf, N. 2004. How learner corpus analysis can contribute to language teaching: A study of support verb constructions. In G. Aston, S. Bernardini and D. Stewart (eds.), *Corpora and language learners.* Amsterdam/Philadelphia: John Benjamins. pp. 109-124.

Nesselhauf, N. 2005. *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Nizonkiza, D. and K. van de Poel. 2014. Teachability of collocations: The role of word frequency counts. *Southern African Linguistics and Applied Language Studies* 32: 301-316. doi:10.2989/16073614.2014.997061

Nizonkiza, D., T. van Dyk and H. Louw. 2013. First-year university students' productive knowledge of collocations. *Stellenbosch Papers in Linguistics Plus* 42: 165-181. doi:10.5842/42-0-143

Pawley, A. and F.H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J.C. Richards and R.W. Schmidt (eds.), *Language and communication*. London, New York: Longman. pp. 191-226.

Rayson's Log-likelihood calculator. Available online: http://ucrel.lancs.ac.uk/llwizard.html. (Accessed 3 January 2014).

Richards, J.C. and R. Schmidt. 2002. *Longman dictionary of language teaching and applied linguistics*. London: Longman.

Ronan, P. and G. Schneider. 2015. Determining light verb constructions in contemporary British and Irish English. *International Journal of Corpus Linguistics* 20(3): 326-354. doi:10.1075/ijcl.20.3.03ron

Shin, D. and P. Nation. 2008. Beyond single words: The most frequent collocations in spoken English. *ELT Journal* 62(4): 339-348. doi:10.1093/elt/ccm091

Scheepers, R.A. 2014. *Lexical levels and formulaic language: An exploration of undergraduate students' vocabulary and written production of delexical multiword units*. Doctoral dissertation. University of South Africa.

Scheepers, R.A. 2016. The importance of vocabulary at tertiary level. *Journal for Language Teaching,* 50(1): 53-77. doi:10.4314/jlt.v50i1.3

Schmitt, N. and R. Carter. 2004. Formulaic sequences in action: An introduction. In N. Schmitt (ed.)*, Formulaic sequences*. Amsterdam/Philadelphia: John Benjamins. pp. 1-22.

Scott, M. 2008. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Shirato, J. and P. Stapleton. 2007. Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research* 11(4): 393-412. doi:10.1177/1362168807080960

Simpson-Vlach, R. and N.C. Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31(4): 487-512. doi:10.1093/applin/amp058

Sinclair, J.M. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stein, G. 1991. The phrasal verb type "to have a look" in modern English. *IRAL* 29(1): 1-29. doi:10.1515/iral.1991.29.1.1

Stubbs, M. 2001. *Words and phrases: Corpus studies of lexical semantics*. Oxford/Massachusetts: Blackwell.

Van Rooy, B. 2006. The extension of the progressive aspect in Black South African English. *World Englishes* 25(1): 37-64. doi:10.1111/j.0083-2919.2006.00446.x

Van Rooy, B. 2011. A principled distinction between error and conventualized innovation in African Englishes. In J. Mukherjee and M. Hundt (eds.) *Exploring second-language varieties in English and learner Englishes: Bridging a paradigm gap*. Amsterdam/Philadelphia: John Benjamins. pp. 189-208.

Van Rooy, B. 2013. Corpus linguistic work on Black South African English. *English Today* 29: 10-15. doi:10.1017/s0266078412000466

Viberg, A. 1996. Basic verbs in second language acquisition. *Revue française de linguistique appliquée* 7: 61-79.

Wang, Y. and P. Shaw. 2008. Transfer and universality: Collocations used in advanced Chinese and Swedish learner English. *ICAME Journal* 32: 201-232.

Wierzbicka, A. 1982. Why can you *have a drink* when you can't **have an eat*? *Language* 58(4): 753-799. doi:10.2307/413956

Wittenberg, E. and M.M. Piñango. 2011. Processing light verb constructions. *The Mental Lexicon* 6(3): 393-413. doi:10.1075/ml.6.3.03wit

Wray, A. 2000. Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics* 21(4): 463-489. doi:10.1093/applin/21.4.463

Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Yan, Q. 2006. A corpus-based analysis of the verb "do" used by Chinese learners of English. *CELEA Journal* 29(6): 37-41. Available online: http://www.celea.org.cn/teic/70/70-37.pdf (Accessed 12 March 2012).