Stellenbosch Papers in Linguistics, Vol. 47, 2017, 133-152
doi: 10.5842/47-0-266

Students' use of academic vocabulary in comparison to that of published writers: A corpus-driven analysis

Trish Cooper

Wits Language School, University of the Witwatersrand, South Africa
E-mail: Trish.Cooper@wits.ac.za

Abstract
An aspect of vocabulary research that tends to be somewhat neglected is that based on qualitative investigation. While a number of studies have considered the differences in vocabulary size between first-language (L1) and additional language (AL) speakers of English, there has been relatively little in-depth investigation into the nature of the vocabulary differences between these groups. The aim of this paper is to shed light on some of the vocabulary features of both L1 and AL student writing in relation to published writing as a benchmark. This study is based on the results of a qualitative investigation conducted using a corpus-driven approach which focused on differences in the use of academic vocabulary by both L1 and AL groups across first-, second- and third-year psychology students. The method used to identify vocabulary differences was keyness analysis, in which vocabulary items are compared on the basis of significantly different frequencies. One of the patterns that emerged serves to support the assumption that L1 students have a better grasp of academic vocabulary than AL students, as there are a greater number of grammatical, semantic and collocational idiosyncrasies in AL writing. The analysis also confirms that high achievers tend to use a broader range of academic words than low achievers. Given the evidence that a good knowledge of academic vocabulary in particular is essential for success at the level of tertiary education, the results of this study contribute to the question of what the specific vocabulary needs of undergraduate students are within the university context.

Keywords: academic vocabulary, first- and additional-language speakers, student writing, corpus analysis, qualitative study, keyness

1. Introduction

An aspect of vocabulary research that tends to be somewhat neglected is that based on qualitative investigation. While a number of studies have considered the differences in vocabulary size between first-language (L1) and additional language (AL) speakers of English, there has been relatively little in-depth investigation into the nature of the vocabulary differences between these groups. The aim of this paper is to shed light on some of the vocabulary features of both L1 and AL student writing in relation to published writing as a benchmark. The argument for regarding the published corpus as the benchmark is based on the

grounds that "the Expert corpus is regarded as a model of the sort of academic writing to which undergraduates … should aspire" (Scheepers 2014:97) and so serves as point of reference for an accepted standard of academic writing. This paper is based on the results of a qualitative investigation conducted using a corpus-driven approach which focused on differences in the use of academic vocabulary by both L1 and AL groups across first-, second- and third-year psychology students.

The principal assumption underlying the research conducted on vocabulary in this study is that there is a difference in the way in which various groups of students use academic words. A detailed, qualitative analysis of select academic words such as *instance, hence, job* and *specific* serves to illustrate how these words are used in context by students of diverse language backgrounds and varying academic levels. Although such qualitative analyses can only highlight a few selected aspects of language use as a result of the in-depth nature of the corpus-based approach, it is hoped that in doing so they serve to demonstrate some of the merits of corpus linguistic investigation at a micro-level. This level of investigation allows for scrutiny of every occurrence of the word being considered, and so reveals syntactic, semantic and pragmatic features of the word as it is used in a range of contexts.

2. Background to the study

This study was conducted at Wits University in Johannesburg and is based on the development of a corpus of student assignments (approximately 3.6 million words) and related journal articles (approximately 1.8 million) – a corpus of over 5.4 million words in total. As the participants were registered for an undergraduate degree, majoring in psychology, the articles were drawn from psychology journals and selected on the basis that they were linked to the assignment topics. The collection of assignments that comprised the student corpus thus serves to track the students' writing over the three years of their undergraduate degree, from first year to third year.

As the focus of this paper is on academic vocabulary, it is necessary to describe the categorisation of these items within a framework of different types of vocabulary and explain the relationship between academic vocabulary and other categories. Academic vocabulary is generally regarded as one of four sets of vocabulary items classified according to frequency of occurrence and distribution (Coxhead and Nation 2001:252; Hyland and Tse 2007:236). The first set comprises high-frequency words that consist of about 2,000 word families and which generally cover between 70% and 80% of texts. The second set is academic vocabulary, some of the most frequent examples of which were compiled into a list of 570 words that cover about 8.5% to 10% of the running words in academic texts (Coxhead 2000). The third set is technical vocabulary, which is subject-specific and "which provides coverage of up to 5% in a text" (Coxhead and Nation 2001:252). In contrast to academic vocabulary, technical vocabulary "occurs in a specialist domain and is part of a system of subject knowledge" (Chung and Nation 2004:252). The last set is that of low-frequency vocabulary items which occur infrequently and are usually restricted in range.

The principle on which lists of academic vocabulary have been developed is that a fairly wide range of words, not commonly found in non-academic texts, occurs regularly throughout academic texts across all disciplines. The features of academic vocabulary are then that these items:

> (a) [are] reasonably frequent in most academic texts from a wide range of academic disciplines,
> (b) [are] relatively infrequent in other types of texts such as novels or colloquial spoken texts,
> (c) [come] largely from French, Latin or Greek, and (d) [are] not obviously connected with any one subject area.
>
> Wang Ming-Tzu and Nation (2004:292)

Words such as *analyse*, *criteria*, *establish*, *proportion* and *subsequent* (Coxhead 1998) play a supportive rather than a central role and so are not highly salient (Coxhead 2000). In terms of this perspective, academic vocabulary items, unlike technical terms, would not necessarily strike the reader as words that require special attention. This view ties in with Durrant's definition of academic vocabulary as:

> sub-technical words which are common across academic disciplines, but which may cause problems for learners because they are neither sufficiently frequent in the language as a whole to be learnt implicitly nor part of the technical lexicon which is likely to be taught as part of subject courses.
>
> Durrant (2009:157)

It seems obvious that vocabulary knowledge should contribute to reading comprehension, since one of the key requirements for comprehension is that the reader has an understanding of the words in the text. In support of this assumption, evidence confirming the link between vocabulary knowledge and reading comprehension has been provided by a number of studies (Clark and Ishida 2005; Cobb and Horst 2001; Stæhr 2008). Without sufficient vocabulary to interpret the meaning of unfamiliar words, L2 learners are unable to reach the comprehension threshold required to understand texts. This is supported by Cooper's (1999:88) finding that, in a study of the relationship between vocabulary and academic performance, academic vocabulary was the most significant indicator of academic performance, as 45% of students who failed the academic vocabulary test failed the year. In this study of vocabulary size among undergraduate students, Cooper (1999, 2000b) conducted receptive vocabulary tests based on three levels of frequency: the 1,000 and 2,000 word lists, the University Word List (Nation 1990) and the advanced word list. An analysis of the results showed that the strongest correlation existed between academic vocabulary scores and academic performance, as measured by the students' examination results. Cooper concludes that a considerable proportion of these L2 degree students do not have the academic vocabulary required to meet the lexical demands of the reading material on which their studies are based. It is evident that these students require explicit vocabulary instruction if they are to attain the lexical threshold necessary for undergraduate study. Similarly, Santos (2004) states that knowledge of academic vocabulary has been found to distinguish academically well-prepared from under-prepared learners, regardless of background.

Given the evidence that a good knowledge of academic vocabulary in particular is essential for success at the level of tertiary education, it is anticipated that the results of this qualitative study may contribute to the question of what the specific vocabulary needs of undergraduate students are within the university context.

3. Analysis of academic vocabulary in the student corpora

For the purposes of this study, an assessment of differences in the use of academic vocabulary by diverse student groups was conducted through qualitative investigation. The first of these groups was L1 versus AL speakers, while the second group was the high- versus low-achieving students. A specific area of interest within the study of vocabulary is the extent to which academic vocabulary items occur in isolation, their meaning independent of other words, or whether they are commonly linked either to function words or to nouns, verbs, adjectives or adverbs by means of a collocational relationship. These relationships are examined by means of keyness analysis, in which vocabulary items are compared on the basis of significantly different frequencies. The strength of association between keyword and collocate was then measured by means of log-likelihood scores. Log-likelihood is "a measure of error, or unexplained variation, in categorical models, based on summing the probabilities associated with the predicted and actual outcomes" (Field 2005:736, 221). An algorithm "converts the difference between the two (scores) into a number which indicates the strength of the collocation – the higher the number, the stronger the collocation" (Baker 2006:101).[1]

Evidence of academic collocations was sought in the corpus of psychology journal articles, and the results were compared with the occurrence of similar collocations in the corpus of student writing (Cooper 2016). The aim of this comparison was to determine the degree to which students are able to emulate the use of collocations typical of published writing in the academic context and so whether students require explicit guidance on common collocational relationships between academic vocabulary items and other words.

As a first step in this qualitative analysis, a measure of keyness was employed to determine which of the academic words in the student corpus differed most significantly from those in the published corpus. The computer program used for identifying keyness within the corpus was WordSmith Tools (WST), version 6.0 (Scott 2012), a program able to analyse lexical bundles or word clusters, collocates, word frequencies and keywords. The WST program identifies 'keyness' in lexical bundles as those that have significantly different densities of occurrence in two selected corpora. Keyness is therefore a measure of saliency rather than frequency, as it takes into account which words occur significantly more often in one text than in another (Baker 2006:125). In the sense that it measures the use of vocabulary in one corpus against the use of that vocabulary in another corpus, keyness represents the degree of overuse and underuse of certain vocabulary items within one corpus in relation to the other. In this case, as the benchmark is the language used by published writers in journal articles, the concepts of 'overuse' and 'underuse' are defined in terms of what is considered 'appropriate' use within an academic context. A feature is regarded as being overused when it is used significantly more relative to the reference corpus and underused when it has significantly fewer occurrences than in the reference corpus. The interest in the disproportionate use of particular vocabulary items is based on the view that compliance with the accepted norms of academic writing is rewarded.

[1] Paul Rayson (n.d.) provides a scale in terms of which the significance of log-likelihood values can be interpreted:

- 95th percentile; 5% level; $p < 0.05$; critical value = 3.84
- 99th percentile; 1% level; $p < 0.01$; critical value = 6.63
- 99.9th percentile; 0.1% level; $p < 0.001$; critical value = 10.83
- 99.99th percentile; 0.01% level; $p < 0.0001$; critical value = 15.13

The tendency of non-native language speakers to overuse certain lexical items, referred to by Hasselgren (1994) as "lexical teddy bears", and underuse others has been reported on by a number of researchers (Ädel and Erman 2012, Chen and Baker 2010, De Cock 2000, Granger and Paquot 2009, Hasselgren 1994, Paquot 2010). The general conclusion is that "learner usage tends to amplify the high frequencies and diminish the low ones" (Lorenz 1999 cited in Paquot 2010:143).

As 'appropriateness' was defined in this context as an assessment of the 'standard' use of vocabulary, it was operationalised in terms of accuracy of usage in relation to the use by published writers. Investigation into this aspect then required in-depth analysis of 'standard' as well as 'non-standard' lexical bundles by means of concordance lines. Part of the qualitative aspect of this study is thus the investigation of concordance lines using WST (2012). A concordance is "a list of all the occurrences of a particular search term in a corpus, presented within the context that they occur in; usually a few terms to the left and right of the search term" (Baker 2006:71). The search term may be a phrase, as demonstrated in the examples of concordance lines below. (1) shows the concordance lines from the published corpus for *on the basis of* : (1a) are the L1 concordances and (1b) the R1 concordances.

(1) a. to create synergies that are beyond those *attainable* **on the basis of** a more homogeneous input. Hence,

research purposes these participants were *chosen* **on the basis of** their established relationship with the

b. status in terms of individual accomplishments, rather than **on the basis of** *ascribed* attributes (sex, age, family name

cultural heritage. Future researchers should create groups **on the basis of** *careful* assessment of cultural orientation.

The WST program presents two levels of concordance analysis: the first lists every occurrence of the search term within the context of the sentence in which it occurs, providing a segment of the sentence as well as of any adjacent sentence, with the number of words given depending on the number of characters specified for the concordance line. The examples in Figure 1 illustrate that the words on either side of the search term (given in bold) can be sorted alphabetically – either immediately to the left of the search term (L1) or to the right (R1), as indicated by italics. The second level of concordance analysis provides access to the text in which the search term occurs. Concordances are therefore made possible by the application of specific software such as the WST program, and form an essential tool within corpus linguistics as they provide an effective means of analysing the context in which particular items occur.

It should be noted that the overarching term used in this study to refer to 'non-standard' forms is the more politically neutral 'idiosyncratic'. Van der Walt and Van Rooy (2002:114) point out that "the word 'standards' has been a sensitive and controversial one … in South Africa" as the norms of Black South African English (BSAfE) have yet to be established. The sensitivity around the use of the word 'standard' relates to the inherent elitism and discrimination in regarding the 'accepted standard' as that closest to standard British English (Webb, 1996, cited in Van der Walt and Van Rooy, 2002). For this reason, any variations from the standard must

be identified with caution, and the distinction clearly drawn between error and "conventionalised innovation" (Van Rooy, 2011). It should be noted, however, that the terms 'non-standard' and 'error' continue to be used in the linguistic analysis of both New and Learner Varieties based on factors such as frequency and stability.

4. Comparison of academic vocabulary in L1 and AL corpora

The first comparison to be conducted was based on the academic vocabulary particular to the additional language students, with the first-language corpus serving as the reference corpus. The results are presented in Table 1 below.

Table 1[2]: Academic words in the AL student corpus with significantly different frequencies from those in the L1 corpus

Key word	Freq. in core corpus	% in core corpus	Normalised rate in core corpus	Freq. in ref. corpus	% in ref. corpus	Normalised rate in ref. corpus	Keyness value*
instance	430	0.04	36.45	198	0.01	10.42	233.68
hence	297	0.03	25.18	203	0.01	10.68	90.70
job	1 299	0.11	110.12	1 459	0.08	76.80	88.22
subordinates	103	0.009	8.73	49	0.003	2.58	53.89
code	105	0.009	8.90	69	0.004	3.63	34.44
evident	284	0.02	24.07	684	0.04	36.01	*-34.19*
occurs	495	0.04	41.96	1 099	0.06	57.85	*-36.54*
individuals	1 538	0.13	130.38	2 990	0.16	157.39	*-36.75*
awareness	136	0.01	11.53	391	0.02	20.58	*-36.85*
occur	378	0.03	32.04	886	0.05	46.64	*-39.08*
previously	77	0.006	6.53	313	0.02	16.48	*-62.61*
community	1 195	0.10	101.30	2 544	0.13	133.92	*-65.40*
specific	504	0.04	42.72	1 262	0.07	66.43	*-74.32*
individual	2 891	0.25	245.08	5 725	0.30	301.36	*-84.09*

*Significant at $p < .0001$

This table provides both positive and negative keyness values for the academic vocabulary in the AL corpus. While the positive values represent those words that occur more frequently in the AL corpus than in the L1 corpus, that is, those that are overused relative to the L1 corpus, the negative values (in italics) represent results of the inverse relationship, that is, those words that occur more frequently in the L1 corpus than in the AL corpus, and so are underused in the AL corpus. As a means of delimiting the scope of this study, in light of the fact that the keyness threshold for all of these academic words is higher than $p < .0001$, the focus of this analysis is restricted to those items that reflect substantial differences in the patterns of usage between the two corpora being compared.

The main methods applied in the course of this qualitative analysis were the investigation of concordance lines, as illustrated in Figure 2 below, and the study of short extracts from the

[2] KEY to Table 1: Normalised rate calculated per 100 000 words; core corpus – AL (1 179 619 tokens); ref. = reference corpus – L1 (1 899 689 tokens)

students' essays based on extensions of the concordance lines, both of which serve to provide examples of actual use by the students. In each analysis, the primary focus was on the collocations in the immediate context of the key word. The first set of concordances to be considered focuses on the key word *instance* which has the highest keyness value in the AL corpus relative to the L1 corpus. A detailed investigation into the use of the word *instance* in both AL and L1 corpora reveals the following collocates:

(2) a. of marriage and in the absence of a father figure. In *this* **instance** children will take on their maternal last name,

b. as her need for self-regard was slowly being met. A *key* **instance** in which we see that Precious had finally started

c. things such as stealing. About three of them have *had* **instance** of excessive coughing though it was when they

d. , 16 year old from a dysfunctional family does to a *large* **instance** perpetuate negative behaviour and actions but

e. virtually interacting on their smartphones and at the *same* **instance** driving a car. However, there are positive and

f. they already know, from a previous instant with the *similar* **instance**. This may increase someone's level of resilience

g. by influences within the environment. In *some* **instance** one is not shaped by society, a person can

h. detected by an authoritative figure, further exacerbated *the* **instance** of second hand smoking. This meant that the

The range of examples presented in these concordance lines serves to illustrate two correct uses of the key word *instance*, and a number of uses that are idiosyncratic. The phrase *in this instance* (2a) within the context of this sentence refers to a particular case, and so conforms to the generally accepted meaning of the word. The second example of correct use is provided in the phrase *a key instance* (2b). While a fairly uncommon collocation (LL: 10.70), this is both grammatically and semantically correct, and reflects a good understanding of the words *key* and *instance*.

As the focus of this study is on grammatically, semantically and collocationally idiosyncratic uses, the remaining examples illustrate the idiosyncratic use of *instance* both grammatically and semantically. Examples (2c) and (2g) represent the only cases in which the error is grammatical, as *instance* is used in the singular form rather than in the plural (*have had instances* and *in some instances*). The phrase *large instance* in example (2d) is an idiosyncratic use (LL: 8.49) which contrasts with the more standard adjectives preceding *instance* in the L1 corpus, such as *first* (LL: 31.65), thereby creating an unusual collocation. The error in example (2e) derives from apparent confusion of *instance* with *instant* ('a precise point in time'),

although both words are incorrect in this context as the phrase ‘at the same time’ would more appropriately express the intended meaning.

The broader context in which example (2f) occurs is presented below to assist in the construal of the intended meaning.

(2) f. This also means that children have the ability to think thoroughly about the way in which they are being responded to in a particular incident, so, instead of not complying with the bully and end up being hurt, they would rather avoid that by giving the bully what they want as they already know, from a previous instant with the similar instance.

This example again stems from confusion between *instant* and *instance*, although in this case the student has juxtaposed both forms in an attempt, it would seem, to express the idea either of *a previous instance* or, alternatively, *a previous incident*.

The last example in the list of concordances above (2h) appears to be based on confusion between the words *incident* and *instance*, as the sentence would be more easily comprehensible were *instance* to be replaced by *incidents*. This is illustrated by the full extract from which the concordance line is drawn:

(2) h. The fact that the smoking girls would secure and lock themselves within the toilet stalls for fear of being detected by an authoritative figure, further exacerbated the instance of second hand smoking.

It has been recognised that words with similar orthographical and/or phonological forms exacerbate the difficulties of learning new words and so are often confused in the language acquisition process (Laufer 1990, Nation 2001, Schmitt 2010).

This detailed analysis of the various uses of the academic item *instance* provides some insight into what are likely to be typical problems experienced by AL students in applying it correctly when writing. In addition to problems of incorrect use, the AL students are clearly overusing this form, given the comparative normalised rates of occurrence (Table 1).

The next academic item to be considered from the AL student corpus is *hence*, which has the second highest keyness difference. This overuse is illustrated in the following extracts from the AL student essays:

(3) a. With that said it may be argued that it is hard to look at the self as solely private because of the interaction that exists as people are growing, hence being socialised. Therefore this helps in understanding that an individual may also be a subject of ideology used to form their identity, hence in South Africa, class and race were used together to formulate black and white identities … .

b. Attention is a process of concentrating of specific features of the environment or on certain thoughts or activities. Hence when one speed reads they include selective attention which is when they exclude of other features of the environment, meaning that they keep focus to what they are reading and ignore distraction, hence this is done through the limited attention as well which is the in capacity and timing.

It is evident from extract (3a) that *hence* is intended to convey the sense of the adverb *thereby* ('thereby being socialised') and the prepositional phrase *in this way* ('In this way, in South Africa, …'). The first use of *hence* in extract (3b) has the meaning *therefore* ('therefore when one speed reads'), and so is an example of the appropriate use of this word. However, the second use of *hence* in this extract is redundant and should be deleted, as it does not add to the sense of cohesion or structure of the argument. Further examples of this are provided in the extracts below:

(3) c. **Since** its capacity is assumed to be limited hence a speed advantage could interact with the delay of information from working memory since less of the proceeding information would decay simply the reason being that of the passage of time.

d. According to Durand individual with anorexia successful lose weight, **however** hence their fear of being obese or gaining weight encourage them to workout in order to maintain weight.

e. **Due to** roles that were associated by men, hence they were views the specie that experience low stress and strain even if there are exposed to stressful situations.

f. Skill variety is another factor that is not part of the job because the skills needed and used in this job are limited hence **that is why** the interviewees feel that their job is not significant.

In the case of (3c), (3d) and (3e), the conjunction *since*, the adverb *however* and the adjective *due to*, respectively, which occur prior to *hence*, make the use of the adverb unnecessary in that it does not add to the meaning being conveyed. Similarly, the phrase *that is why* in (3f) makes the use of *hence* redundant as it conveys the same sense. These unnecessary insertions of the word *hence* may be regarded as semantic errors, as they seem to reflect a poor understanding of the word's meaning and function within the context of each example.

The next set of examples serves to illustrate grammatical idiosyncrasies in the use of *hence*. It is clear from the analysis of concordance patterns in both the L1 and published corpora that *hence* is typically followed by a noun phrase and less commonly by a prepositional phrase or a verb phrase. The ten most frequent R1 collocates in the L1 corpus are *the, it, they, we, this, there, perceptibility, in, women* and *these*. While there is a considerable amount of overlap between the most frequently occurring R1 collocates in both AL and L1 corpora, it is apparent that one of the sources of grammatical error in the AL corpus is the omission of either the subject or object following the use of the adverb *hence*, as illustrated in the concordance lines in (4).

(3) g. and it does differ due to the ratings of high levels of crime. **Hence** *can* be argued that it is the same crime and it

h. situations in which the individual has to step up and act. **Hence** *said* earlier that in the process of developing as a

i. their life until they achieve these goas (in adulthood) **hence** *why* the increase of these traits starts to become

j. become more consistent yet retain the potential for age. **Hence** *why* it was mentioned in the introduction that

This analysis suggests that many AL students appear to have only a vague understanding of the grammatical patterns and meaning of the word *hence*, with the result that it is often not used appropriately or in the correct context. The extracts from student essays below illustrate this point. In each example, the word or phrase that would more appropriately replace *hence* is given in small capital letters at the end of the extract and underlined.

(3) k. The authority figure explains why their procedure is good and the reasons why it might be more effective hence resulting in cognitive change in those influenced. THEREBY

l. A highlighted above Bud is desperate to make money therefore he employs reward political tactics in order to charm people hence get his way in. AND SO/ THEREBY

m. Thus it can be argued that speed readers are brilliant readers when they read some kinds of material for some purpose, hence when reading other kinds of materials for other purposes there is no relationship between speed of reading and the ability to comprehend. ALTHOUGH

It may be concluded that two of the more common mistakes being made by AL students in the use of *hence* result from confusion between *hence* and *thereby* and from the overuse of *hence* where it should be omitted, as it is either unnecessary in the context or the meaning is adequately conveyed by an alternative adverb, adjective or conjunctive.

The next set of concordances to be considered focuses on the keyword *job*, which has the third highest keyness value. A predominant feature within the AL corpus with regard to *job* is its repeated use within single paragraphs, suggesting overuse by the AL students. This pattern occurs in a number of student texts, as illustrated in the extract given below (4a).

(4) a. Holman suggests that job design has to do with structuring of the actual features of a job itself. This entails the amount of time, the type of job, what skills are required and how difficult the job is. Redesign in the light speaks of the elements that need to be restructured in order to bring about organisational effectiveness. This also includes the design of where and when the job is done. Included in the description of job redesign techniques are job enlargement, job simplification, job enrichment, and job rotation. Studies conducted under the mediating role of

> job characteristics in job redesign interventions conclude "that the effects of employee participation in job redesign on well-being are a result of changes in job characteristics rather than participation in change per se".

While the academic item *job* is key to this essay in which students were required to analyse various aspects of a job of their choice and make recommendations for improvements, there is little evidence that the L1 students repeated the key word *job* to the same extent as is evident in the AL corpus. It is clear that these students could have avoided such repetition through the use of pronouns, synonyms such as *work* where appropriate, and omission. The AL corpus example below illustrates these points in changes made to an excerpt from a student essay.

(4) b. The job specific's entails aspects ~~of the job~~ that are separate from the organisation in which the individual works in. The specific job that this essay will explain in-depth is that of two bus drivers at Regina high school who preferred to remain anonymous. Furthermore, in order to construct a detailed ~~job~~ analysis I made sure that the two participants were from the same hierarchical ranking within the organisation, because that would enable me to gain information about the job itself rather than the personal opinions ~~about the job~~ of the employees.

This exercise illustrates that AL students would not only benefit from workshops in which the appropriate use of cohesion markers in general is clarified, but would also benefit from strategies on how to avoid unnecessary repetition in their writing.

The focus of this analysis now shifts from overuse to underuse by the AL students in relation to the L1 speakers. Of the nine items in Table 1 with a negative keyness value, only those that revealed particularly interesting differences in the way in which the items were used in the AL and L1 corpora are discussed in detail.

Investigation into the use of the word *specific* in the AL corpus revealed two non-standard patterns. The first relates to the phrase *to be specific*, and the second to the phrase *in specific*. The first set of concordance lines provided below suggests that AL students are confusing the adjective *specific* with the adverb *specifically*.

(5) a. feel that as a humanities student, in social work to *be* **specific**, I do relate with the community in a professional

b. of failing to develop resilience, adolescent's period to *be* **specific** is seen as a time of risk. And these risks are

c. the latter may occur particularly in young women to *be* **specific**, and how it may shape youth identity through the

d. evident that people and other species, chimpanzees to *be* **specific** use tools to carry out tasks hence form culture.

e. to discuss the identity development and violent crime *using* **specific** Erikson's theory of psychosocial theory.

It is evident that replacing the phrase *to be specific* with the adverb *specifically* would make the interpretation of these extracts easier, although (5a) and (5d) also require a change in the word order, with *specifically* occurring before the prepositional phrase "in social work" in (5a) and before the noun "chimpanzees" in (5d). In addition to these grammatical errors involving parts of speech, it is clear from both (5c) and (5e) that the use of *specifically* would in fact be redundant. In the case of (5c), it is preceded by *particularly*, which expresses the same meaning, and in (5e) it is superfluous and can be omitted without any loss of meaning. The examples above from the AL corpus contrast clearly with those in (5f) and (5g) from the L1 corpus, as these represent more standard uses of the phrase *to be specific*.

(5) f. Therefore, eating disorders in South Africa appear to *be* **specific** to adolescent females within an urban

g. an individual's life course development, there appear to *be* **specific** indicators that need to be established in order for

A further point of interest in relation to the AL students' use of the copula verb *be* immediately before *specific* is that, while it is clearly used to express conditionals in the published corpus, there is no evidence of this use in the AL corpus. The examples below, taken from the published corpus, serve to illustrate this point, as the phrase *be specific* is preceded by a modal expressing possibility.

(5) h. the addressees in both conditions. The behaviors might *be* **specific** to teaching children. Against this interpretation is

i. argument earlier that results from forced choice could *be* **specific** to that method. If we are to disregard results

j. a caveat in interpreting our data is that this effect may *be* **specific** to our experimental conditions, in that we

The second idiosyncratic pattern of use relating to *specific* is evident in the following concordance lines from the AL corpus.

(5) k. to say on the subject in general as well as pretend play *in* **specific** attributing to it a role more often in cognitive and

l. family and work roles and lower satisfaction. Women *in* **specific** are faced with numerous challenges when

These examples are similar to those discussed in (5a) to (5e), as they reflect a degree of confusion in that the preposition + adjective *in specific* should be replaced by the adverb *specifically*. Reference to the published corpus clearly illustrates that the phrase *in specific* is typically followed by a plural noun form (4m to 4p).

(5) m. theory designed to predict and explain human behavior *in* **specific** contexts. Because the theory of planned behavior

n. to genetic factors that define propensities to grow *in* **specific** directions at specific ages during the life course.

o. dispositions tend to be poor predictors of behaviour *in* **specific** situations. General attitudes have been assessed

p. GMV. Nonetheless, the presence of structural deficits *in* **specific** regions may be important for the interpretation of

The discrepancy in the use of *specific* by the L1 and AL groups may be due to the AL students' apparent confusion regarding the correct applications of the adjectival form, with the result that they are more likely to avoid using it than are the L1 students.

An analysis of the word with the third highest negative keyness ranking, *community*, as used by the AL and L1 students, shows little real difference between the two corpora, presumably because the word is unambiguous in the context of the third-year essay in which it was used. There is a considerable degree of difference, however, in the frequency of occurrence of the pronoun *my* immediately preceding *community*. This pronoun has a frequency of 2% in the AL corpus, with the collocational pairing having a log-likelihood ratio of 104.16. In the L1 corpus, on the other hand, the pronoun *my* has a frequency of 4%, with a far higher log-likelihood ratio of 1036.5. It may be argued that the far more predominant use of *my* to describe *community* in the L1 corpus reflects the L1 students' greater use of the first person in this essay. This inference is supported to some degree by the collocational links between the pronoun *I* and the key word *community*, which were stronger in the L1 corpus (LL: 431.82) than in the AL corpus (LL: 308.47). It is possible that, although the essay topic encouraged use of the first person, the AL students are less confident about their writing skills than the L1 students, and so more reluctant to move away from the more conventional use of the third person. This interpretation is supported by Hyland's argument regarding the writer's stance or position in student assignments. Hyland (2002:1091) proposes that the reluctance to use the first person in academic writing is particularly problematic for L2 writers, as "the individualistic identity implied in the use of *I*" frequently runs counter to the representations of self inherent in their own cultures. In subsequent research, Hyland (2012:66) found that "in a corpus of research articles … half the occurrences of *I* collocated with the presentation of arguments or claims, while this was the least frequent use in undergraduate reports, where writers were reluctant to make such strong personal commitments and instead mainly used *I* to state a purpose". As this line of conjecture raises a number of questions, the density of reference to first person as opposed to third person and the use of agent-evacuated passives could be a matter for future research.

As there is little evidence of difference between the corpora in the use of the adverb *previously*, the next item to be examined is *occur*, which has the fifth-highest negative keyness value. While this verb is used similarly in both corpora there is a substantial difference in the density of modals before the verb. In support of this claim, the following table shows the log-likelihood values for a selection of modals which collocate with *occur* in each student corpus:

Table 2: Modal collocations for the key word *occur* in the AL and L1 corpora

Modal	LL value: AL corpus	LL value: L1 corpus
can	293.59	864.07
may	133.14	747.01
will	128.45	575.46
might	26.16	53.62
could	24.70	121.63

These log-likelihood values indicate that there is a far stronger association between the verb *occur* and the preceding modals in the L1 corpus than in the AL corpus. Although *will* expresses intent, the remaining modals indicate possibility or probability, suggesting that the L1 students are more inclined to hedge their arguments, while AL students tend to be more assertive, using fewer qualifiers, as argued by Hyland (1994).

The discussion of *occur* concludes the detailed comparative analysis of select words that have significantly different frequencies in the AL and L1 corpora. Despite the apparently disparate issues discussed in this section, qualitative, corpus-driven comparisons of these fairly substantial AL, L1 and published corpora provide fine-grained analyses that help to explain very specific problems encountered by AL students in using academic words, while at the same time helping to explicate the broader quantitative findings of overuse and underuse. The overall impression gained from this analysis is that the use of these academic words by L1 students more closely approximates that of the published writers than does the use by AL speakers.

In addition to the comparison of L1 and AL speakers, this study also compares the writing of students categorised according to academic performance. The high achievers were those whose results placed them in the group of top-third performers overall, while the low achievers were those whose results placed them in the bottom third overall. The next section is a comparison of the use of specific academic words by the high and low achievers.

5. Comparison of academic vocabulary in high- and low-achieving student groups

As the use of academic words by the high achievers at first-year level did not differ significantly from that of the low achievers, and the differences in the third-year corpora relate to the choice of essay topic, the discussion in this section focuses on words from the second-year corpora.

Only two words in the second-year corpus of low achievers differ significantly in terms of frequency from those used by high achievers, that is, *whereby* and *principle*. Table 2 illustrates that, while *whereby* is overused by low achievers in relation to high achievers, *principle* is underused. These differences are explored in this section.

Table 3[3]: Academic words in the PSY200 low-achievers' (LA) student corpus with significantly different frequencies from those in the PSY200 high-achievers' (HA) corpus

Key word	Freq. in core corpus (LA)	% in core corpus	Normalised rate in core corpus	Freq. in ref. corpus (HA)	% in ref. corpus	Normalised rate in ref. corpus	Keyness value*
whereby	145	0.04	41.85	76	0.02	18.29	36.27
principle	44	0.01	12.70	151	0.04	36.35	*-44.27*

*Significant at $p < .0001$

According to the Cambridge International Dictionary of English (1995), the standard meaning of *whereby* is "by which way or method", while the application of *whereby* to mean "in which" is non-standard. This word is therefore typically used to show an instrumental relationship and may be expected to follow nouns such as *approach*, *effect*, *procedure*, *process*, *means* and *method*. Although these uses are clearly evident in the published corpus and, to a large degree, in the writing of high achievers, there is a considerable amount of latitude in the use of *whereby* within the writing of low achievers, as indicated by the examples below.

(6) a. . In that case, if the child is exposed to domestic *abuse* **whereby** the parents of the child engage in physical or

b she is asked by her teacher. The concept of the self *is* **whereby** the individual notices themselves consciously as

c. of the self as well. This then creates an *atmosphere* **whereby** the experiences that are seen as enhancing the

d. is seen as being the same as many patriarchal *societies* **whereby** men are traditionally allowed to physically punish

e. transition from childhood to adolescence is ongoing *change* **whereby** all individual go through. When individuals

f. had to move between different family members in *cases* **whereby** their mother had passed away and their fathers

g. from their parents. This took place in the post-war *context* **whereby** bereavement and deportation were prevalent

h. discussed above, Precious had reached a point in her *life* **whereby** nothing else mattered except doing anything that

i. . For example this is seeing in the hijacking of *car* **whereby** most whites are the ones who are high jacked.

[3] KEY to Table 3: Normalised rate calculated per 100,000 words; core corpus – PSY200 low achievers (346 437 tokens); ref. = reference corpus – PSY200 high achievers (415 415 tokens)

j. , many South Africans had experienced violent *crimes,* **whereby** they were either victims of violent crimes,

k. only takes place through a system of several processes *of* **whereby** each is concerned with the representation of the

l. classmates and her new teacher at each one teach *one.* **Whereby** she started being interested in sitting in the

These examples illustrate a range of uses, from introducing details (6a and 6b) to meaning 'in which' (6c-6d), 'which' (6e), 'where' (6f), 'when' (6g and 6h), and 'as' or 'since' (6i and 6j). There is also evidence of grammatical errors, as in (6k) where a preposition rather than a noun phrase or a verb phrase is used immediately before *whereby*; and in (6l), where the conjunction is used in sentence-initial position. The extracts below (6m and 6n) clearly reflect the extent to which *whereby* has generally been overused by low achievers, with more suitable alternatives provided in brackets at the end of each extract.

(6) m. Research shows that exposure to violence affect boys and girls behaviour differently, whereby traditionally boys would act out and display the violent acts they are exposed to against others while girls display emotional feelings like being sad and lonely, however recent studies show that the roles seem to have changed whereby the boys seem to also display emotional feelings and girls seem to be more aggressive than usual. (SINCE or Ø, AS)

n. It could be argued that the therapist in the film took this passive approach, whereby she allowed for Precious to take hold of the therapy session. She asked her very minimal question, whereby she expected Precious to respond and do most of the talking. (AS, AND)

The fact that the first instance of *whereby* in (6m) could be omitted completely suggests that students occasionally use it to force a link between sentences in an attempt to create the semblance of an argument. This use provides further evidence that, while *whereby* is perceived as a linker, the real sense of the word does not seem to be clearly understood. Although there is less evidence of the diversity of uses employed by the low achievers in the writing of the high achievers, there is nevertheless a degree of variation in the application of *whereby*, as illustrated by the examples below.

(6) o. , even though its definition has a bit of *contradictory* **whereby** it suggests that change does occur but it does

p. was due to her upbringing and socialized *experiences,* **whereby,** her mother constantly utters bad words to her

q. has severe consequences on the women of that *society,* **whereby** violence is mainly used to uphold male

r. to speak each word to themselves as fast as they *can,* **whereby** still retaining the content of the text. Carver

As with the low achievers, *whereby* is used here to introduce detail (6o), and to mean 'in which' (6p), 'where' (6q) and 'while' (6r). However, the standard use is applied more often than in the low-achiever corpus, as it follows phrases such as *feedback activation*, *eye movement behaviour*, *identity foreclosure*, *positive regard*, *psychosocial development*, *socialised experiences* and *top-down processing*.

In contrast to the positive keyness value of *whereby*, *principle* has a negative keyness value, reflecting a particularly low frequency of occurrence in the low-achiever corpus, where it is used in the sense of a primary, basic idea as well as the head of a school, with the latter use based on the incorrect spelling of *principal*. Similarly, both senses of *principle* are used in the high-achiever corpus, including the misspelling of *principal*. However, the high achievers use terms such as *continuity principle*, *discontinuity principle*, *developmental principle*, *epigenetic principle*, *likelihood principle*, *orthogenetic principle* and *psychological principle*, while the low achievers refer to only *epigenetic* and *psychological principles*. The greater proportion of technical terms evident in the high-achiever corpus was also found to occur in the L1 corpus when compared to the AL corpus. One of the possible inferences to be drawn from this finding is that the high achievers may be more comfortable employing psychology terms in context, as they seem less inclined to avoid such terms in their essays. However, it must be noted that this is a tentative conclusion that could be tested by means of further research into the use of academic vocabulary and technical terms by low achievers at undergraduate level.

6. Conclusion

One of the patterns that emerged in the course of this qualitative analysis served to support the assumption that L1 students have a better grasp of academic vocabulary than AL students, as there is a greater number of grammatical, semantic and collocational idiosyncrasies in AL writing. Examples from the AL corpus include the overuse of certain academic words; the non-standard use of adverbs such as *hence*; unusual collocations linked to academic vocabulary; confusion between parts of speech such as adjectival and adverbial forms, as well as between similar lexical forms; and a restricted range of vocabulary items, which reduces the number of options for collocational pairings. Examples from the L1 corpus, on the other hand, reflect more prolific use of the first person and a greater range of collocations in relation to academic words. This qualitative analysis also confirmed that high achievers tend to use a broader range of academic words than low achievers. Findings such as these illustrate the depth of analysis possible in corpus-based studies that make use of the features provided by software such as WST, which enable the researcher to establish keyness, create concordances based on key words or phrases, and explore aspects of the linguistic environment of the key item such as grammatical and collocational links.

In line with Nation's view that "words are not isolated units of language" (Nation 2001:23), Hancioğlu et al. (2008:463) propose that the key feature of an academic text is "the ways in which certain items 'collocate' and 'colligate', in other words, the ways lexical items co-occur with other lexical and grammatical items". A solid grasp of the relationship between academic items and their collocates appears to be one of the variables distinguishing AL from L1 students as well as students of different academic proficiencies. It may therefore be argued that the identification and in-depth analysis of items that are commonly misused with the view to providing concrete examples of how these words should be used in context would be of considerable benefit. As qualitative analysis of the sort conducted for this study, using a

program such as WST, so clearly reveals patterns of both 'standard' and idiosyncratic use, it brings us that much closer to addressing typical idiosyncrasies in student writing and so to providing guides to academic writing that are more closely based on the professional benchmark.

References

Ädel, A. and B. Erman. 2012. Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes* 31(2): 81-92. doi:10.1016/j.esp.2011.08.004.

Baker, P. 2006. *Using corpora in discourse analysis.* London: Continuum.

Cambridge International Dictionary of English. 1995. Cambridge: Cambridge University Press.

Chen, Y-H. and P. Baker. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology* 14(2): 30-49. Available online: http://llt.msu.edu/vol14num2/chenbaker.pdf (Downloaded 7 April 2015).

Chung, T.M. and P. Nation. 2004. Identifying technical vocabulary. *System* 32: 251-263, doi:10.1016/j.system.2003.11.008.

Clark, M.K. and S. Ishida. 2005. Vocabulary knowledge differences between placed and promoted EAP students. *Journal of English for Academic Purposes* 4: 225-238, doi:10.1016/j.jeap.2004.10.002.

Cobb, T. and M. Horst. 2001. Reading academic English: Carrying learners across the lexical threshold. In J. Flowerdew and M. Peacock (eds.) *Research perspectives on English for academic purposes*. Cambridge: Cambridge University Press. pp. 315-329.

Cooper, P.A. 1999. *Lexis and the undergraduate: Analysing vocabulary needs, proficiencies and problems.* Unpublished MA thesis. Pretoria: University of South Africa.

Cooper, P.A. 2000. The weight of words: Considering vocabulary in literacy. *Innovation* 21: 42-47.

Cooper, P.A. 2016. *Academic vocabulary and lexical bundles in the writing of undergraduate psychology students.* Doctoral dissertation. Pretoria: University of South Africa.

Coxhead, A. 1998. *An Academic Word List.* English Language Institute Occasional Publication No. 18. Wellington: Victoria University of Wellington, School of Linguistics and Applied Language Studies.

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

Coxhead, A. and P. Nation. 2001. The specialised vocabulary of English for academic purposes. In J. Flowerdew and M. Peacock (eds.) *Research perspectives on English for academic purposes.* Cambridge: Cambridge University Press. pp. 252-267.

De Cock, S. 2000. Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair and M. Hundt (eds.) *Corpus linguistics and linguistic theory. Papers from the Twentieth International Conference on English language research on computerised corpora (ICAME 20)*. Amsterdam: Rodopi. pp. 51-68.

Durrant, P. 2009. Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes* 28: 157-169. doi:10.1016/j.esp.2009.02.002.

Field, A. 2005. *Discovering statistics using SPSS.* London: Sage.

Granger, S. and M. Paquot. 2009. Lexical verbs in academic discourse: A corpus-driven study of learner use. In M. Charles, D. Pecorari and S. Hunston (eds.) *Academic writing: At the interface of corpus and discourse.* London: Continuum. pp. 193-214.

Hancioğlu, N., S. Neufeld and J. Eldridge. 2008. Through the looking glass and into the land of lexico-grammar. *English for Specific Purpose*s 27: 459-479. doi:10.1016/j.esp.2008.08.001.

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* 4(2): 237-260. doi:10.1111/j.1473-4192.1994.tb00065.x

Hyland, K. 1994. Hedging in academic writing and EAP textbooks. *English for Specific Purposes* 13(3): 239-256. doi:10.1016/0889-4906(94)90004-3

Hyland, K. 2002. Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics* 34: 1091-1112.

Hyland, K. and P. Tse. 2007. Is there an "Academic Vocabulary"? *TESOL Quarterly* 41(2): 235-253.

Hyland, K. 2012. *Disciplinary identities: Individuality and community in academic discourse.* Cambridge: Cambridge University Press.

Laufer, B. 1990. 'Sequence' and 'order' in the development of L2 lexis: Some evidence from lexical confusions. *Applied Linguistics* 11(3): 281-296. doi:10.1093/applin/11.3.281

Nation, I.S.P. 1990. Teaching and learning vocabulary. New York: Newbury House.

Nation, I.S.P. 2001. *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Paquot, M. 2010. *Academic vocabulary in learner writing: From extraction to analysis.* London: Continuum.

Rayson, P. n.d. *Log-likelihood and effect size calculator*. Available online: http://ucrel.lancs.ac.uk/llwizard.html (Accessed 15 August 2013).

Santos, M.G. 2004. Some findings on the academic vocabulary skills of language-minority community college students. *NCSALL*. Available online: http://www.ncsall.net/index.php@id=175.html (Accessed 15 August 2013).

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

Schmitt, N. 2010. *Researching vocabulary: A vocabulary research manual.* Hampshire: Palgrave Macmillan.

Scott, M. 2012. *WordSmith Tools version 6.* Liverpool: Lexical Analysis Software.

Stæhr, L.S. 2008. Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal* 36(2):139-152. doi:10.1080/09571730802389975.

Van der Walt, J.L. and B. Van Rooy. 2002. Towards a norm in South African Englishes. *World Englishes* 21(1): 113-128.

Van Rooy. B. 2011. A principled distinction between error and conventionalized innovation in African Englishes. In J. Mukherjee and M. Hundt (eds) *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a paradigm gap.* Amsterdam: John Benjamins. pp. 189-207.

Wang Ming-Tzu, K. and P. Nation. 2004. Word meaning in academic English: Homography in the Academic Word List. *Applied Linguistics* 25(3): 291-314.