Acquiring a scholar’s voice: vietnamese students mastering academic vocabulary in thesis writing

Adquirir la voz de un erudito: estudiantes vietnamitas dominan el vocabulario académico en la redacción de tesis

Sandra Gollin-Kies** y Olga D. Lambert***

ABSTRACT

We report preliminary results of a study of academic writing by graduate students writing in English in Vietnam, triangulating quantitative and qualitative data to gain insight into students’ perceptions of their writing and their actual performance. The research involves Vietnamese EFL teachers (n = 62) taking a 2-year MA Linguistics (TESOL) program delivered in hybrid mode. The program, jointly taught by an American private university and a large public university in Vietnam, uses an innovative scaffolded thesis model developed by the researchers. In program entry surveys, students mentioned vocabulary most often as a significant challenge for academic writing. In an attempt to quantify the extent of this perceived deficit, we compared the vocabulary that our students used in early and late drafts of their theses against two reference corpora: the graduate level linguistics subcorpus of the BAWE corpus of student academic writing and COCAA 2010-12. Data from the first drafts of the introduction, discussion and conclusion sections shows that the Vietnamese students compared favorably with both native and non-native students in the BAWE subcorpus in terms of lexical density measured as type to token ratio, but their use of both discipline-specific and “core academic” words was slightly more limited, and they relied more on repetition of the most common words in English. Between drafts, the overall length of writing expanded by approximately 17%, and students collectively added 302 word types, including 25 core academic and 11 discipline specific. The study suggests

** Sandra Gollin-Kies, Ph.D. Benedictine University. Lisle, IL, USA. skies@ben.edu
*** Olga D. Lambert, Ed.D. Benedictine University. Lisle, IL, USA. olambert@ben.edu

Keywords: academic writing, corpus analysis, EFL, vocabulary.
that vocabulary cannot be disentangled completely from other aspects of academic writing such as lexical bundles, syntax, and cohesion, and these require more investigation in advanced students writing in an EFL environment.

RESUMEN
Informamos los resultados preliminares de un estudio de escritura académica de graduados que escriben en inglés en Vietnam, triangulando datos cuantitativos y cualitativos para obtener información sobre las percepciones de los estudiantes sobre su escritura y su rendimiento real. La investigación involucra a maestros vietnamitas de EFL (n = 62) que toman un programa de 2 años de MA Lingüística (TESOL) impartido en modo híbrido. El programa, impartido conjuntamente por una universidad privada estadounidense y una gran universidad pública en Vietnam, utiliza un innovador modelo de tesis con andamios desarrollado por los investigadores. En las encuestas de ingreso al programa, los estudiantes mencionaron el vocabulario con mayor frecuencia como un desafío significativo para la escritura académica. En un intento por cuantificar el alcance de este déficit percibido, comparamos el vocabulario que nuestros estudiantes utilizaron en los borradores tempranos y tardíos de sus tesis con dos corpus de referencia: el subcorpus lingüístico de nivel de posgrado del corpus BAWE de escritura académica estudiantil y COCAA 2010-12. Los datos de los primeros borradores de las secciones de introducción, discusión y conclusión muestran que los estudiantes vietnamitas se compararon favorablemente con los estudiantes nativos y no nativos en el subcorpus BAWE en términos de densidad léxica medida como proporción de tipo a token, pero su uso de palabras específicas de la disciplina y “académicas básicas” fue ligeramente más limitado, y se basaron en la repetición de las palabras más comunes en inglés. Entre los borradores, la longitud total de la escritura se expandió en aproximadamente un 17%, y los estudiantes agregaron colectivamente 302 tipos de palabras, incluidas 25 académicas básicas y 11 específicas de la disciplina. El estudio sugiere que el vocabulario no puede separarse completamente de otros aspectos de la escritura académica, como los paquetes léxicos, la sintaxis y la cohesión, y estos requieren más investigación en estudiantes avanzados que escriben en un entorno EFL.
Introduction

The field of English for Academic Purposes, EAP, has always focused heavily on enhancing non-native students’ performance in academic writing, and teachers have sought to discover more effective pedagogies to help such students address their academic writing needs. While grammar has continued to be of primary concern, an increasing number of studies have asserted the importance of developing academic vocabulary (e.g., Berman & Cheng, 2010; Evans & Green, 2007; Evans & Morrison, 2011). In response to this challenge, wordlists have proliferated for general academic use (e.g., Coxhead, 2000; Gardner & Davies, 2014) and highly specific wordlists have been generated in disciplines as diverse as agriculture, medicine, engineering and applied linguistics (Gholaminejad & Sarab, 2020). Wordlists have also appeared for idioms, discourse markers, and other features of academic writing (EAP Foundation, n.d.).

Despite these burgeoning resources, exploration of the literature on academic writing for English language learners, even those with high proficiency, shows that vocabulary continues to be one of the key concerns of both international students and their instructors (e.g., Hinkel, 2003; Hyland & Tse, 2007; Leedham & Cai, 2013). Berman and Cheng (2010) found “significant negative correlations ... among the NNS graduates between students’ perceptions of their language difficulties and their academic achievement as represented by their GPAs” (p. 37). Furthermore, most studies are based in countries where English is the dominant language, and there is considerably less reporting on the vocabulary issues of second language speakers attempting advanced academic writing in English in their own countries, especially in Asia. This paper attempts to address this gap by exploring productive vocabulary use in a graduate program taught in Vietnam.

The MA Linguistics (TESOL) program

The two-year MA Linguistics (TESOL) program is offered jointly by a small private university in the Midwest region of the United States (henceforth referred to as U.S. partner) and a large public university in Vietnam (henceforth Vietnam partner). The program uses a cohort model, with groups of 20-25 students entering, completing all the courses, and graduating in lockstep. It is delivered in hybrid mode (partially
face-to-face in Vietnam and partially online). Half of the program’s 32 credit hours are taught by Vietnam partner instructors via a mix of face-to-face and online sessions, depending on students’ professional work schedules. The other half of the program is taught by U.S. partner instructors, who visit Vietnam every summer to teach intensively and the rest of the time, offer a mix of synchronous and asynchronous online instruction via the integration of the U.S. partner’s learning management system (LMS) and Zoom. Since the summer of 2020, due to the COVID-19 pandemic, the U.S. portion of the program has been delivered fully online.

As part of the program, students are required to complete a master’s thesis of 35-50 double-spaced pages reporting on an original empirical study they have designed and conducted under the supervision of faculty members from both institutions. The program uses an innovative scaffolded thesis model, where instead of taking one four-credit thesis course in their final semester, students develop their academic writing through four one-credit thesis writing seminars interwoven through the academic program.

Scaffolded pedagogy in the MA Ling (TESOL)

Students examine and analyze models of academic writing using a genre-based approach (Swales, 1990), while practicing writing annotated bibliographies, critiques, and other sub-genres as part of their thesis writing process. They receive intensive feedback via the U.S. partner’s LMS at all stages of their thesis writing, and extensive re-writing takes place. Students are encouraged to develop their academic vocabulary through “noticing” (Ellis, 2006) vocabulary use in target academic genres, and through working with corpora and concordance tools (e.g., COCA word and phrase1).

The four thesis seminars are an especially good example of our innovative scaffolded approach. During the first seminar, students learn about academic writing in the discipline, become familiar with the master’s thesis requirements and format, and practice choosing a specific topic in the field of applied linguistics or TESOL, as well as locating, reading, and evaluating peer-reviewed sources relevant to this topic. The second

1 See: https://www.wordandphrase.info/wap_coca.pdf
The seminar focuses on helping students develop a draft of a thesis proposal as well as acquire additional practice with quantitative and qualitative data analysis. During the third seminar, students finalize their proposals and apply for institutional review board (IRB) approval for their projects, and the fourth and final seminar is dedicated to writing, revising, and editing a complete draft of their master’s thesis.

The MA Linguistics (TESOL) students
All students are trained English teachers with at least 2 years of teaching experience required for admission into the program. They work at public and private schools, language centers that cater to learners of all ages and levels (many specializing in IELTS or TOEIC preparation), and universities. Many also supplement their income as private tutors. Their IELTS scores on admission range between 6.0 and 8.0 (CEFR upper B2-C1), with the mean score of 6.5; all are required to achieve 6.5 before graduation. The first four cohorts (n=62 students) had graduated as of December 2021.

Student entrance survey as partial motivation for the study of productive vocabulary
Beginning with Cohort 3, an entrance survey, distributed via a Qualtrics link during orientation, was administered to all students. We quickly noticed that many students identified academic writing, and specifically vocabulary, as a major area of difficulty for them—a finding that served as a partial impetus for this study. In the methods and findings sections, we will report on the results of the entrance survey from Cohorts 3 and 4 (n = 38) that provided insight into how students evaluated their academic writing skills at the start of the program and what their previous experiences with academic writing had been. The survey also identified what they thought were their biggest challenges when it came to academic writing. Our quantitative analysis of the students’ writing sheds light on the accuracy of some of these self-evaluations by students.

Literature Review
The development of large electronic corpora and concordancing software that can be used to investigate word frequencies and distributions in texts has enabled a more sophisticated understanding of the nature and role of
vocabulary in academic writing. At the same time, there is much theoretical
discussion about what constitutes an adequate vocabulary for academic purposes, and researchers differ on how students can best achieve it.

At its most basic level, the amount of vocabulary in a text can be measured in terms of the number of distinct word forms or types, or alternatively as lemmas (a headword and all its inflected and reduced forms), or as word families consisting of a headword, along with its inflected forms and related derived forms. According to Nation (2013), educated adult native speakers of English know around 20,000 word families, but only 6-9,000 word families are needed to comprehend 98% of most written material. However, the vocabulary needed for productive use could be considerably smaller, depending on the text type and purpose (Malmström et al., 2018).

Analysis of large corpora has steadily refined understanding of the specific words students need in order to be competent users of the target language. Since West (1953) developed the General Service List (GSL) of the most common 2,000 words in English, the GSL been updated to the New GSL (Browne et al., 2013a) recognizing that common words from the 1950s are no longer in the top 2,000 words, and new words have taken their place. There has also been continuing research into vocabulary specific to academic reading and writing, with foundational work by Xue and Nation (1984) superseded by Coxhead’s Academic Wordlist (AWL) (2000). Notable recent developments include the New Academic Wordlist (NAWL) (Browne et al., 2013b), and the Academic Vocabulary List (AVL) (Gardner & Davies, 2014), which is based on COCAA, a large academic subcorpus of the Corpus of Contemporary American English (Davies, 2008).

The importance of developing academic vocabulary lists for EAP pedagogy has been vigorously debated (Durrant, 2016; Gardner & Davies, 2014; Hyland & Tse, 2007). Many researchers have argued that general academic wordlists like the AVL do not sufficiently target the needs of students in particular disciplines (e.g., Durrant, 2016; Hyland & Tse, 2007; Martínez et al., 2009). In response, many specialized wordlists have been produced for particular disciplines. Another concern

---

2 See the EAP Foundation website for a useful list and comparison of different wordlists.
is that wordlists based on articles by published professional researchers are unnecessarily extensive for English language learners and contain many words students are unlikely to ever use in their own writing (Durrant, 2016; Martínez et al., 2009).

Broadly speaking, productive vocabulary is used to express meaning through speaking or writing, while receptive vocabulary refers to that used to retrieve meaning from listening or reading, though the relationship between the two types is not clear cut (Nation, 2013; Nizonkiza, 2016). Clearly, one needs to be able to access the vocabulary of the field of study in order to write academically, and receptive vocabulary can be more than 50% larger than productive (Nizonkiza, 2016). Malmström et al. (2018) noted that there is a paucity of research on productive vocabulary for academic writing compared with receptive vocabulary. These researchers compared the use of AVL list words by the mostly native speaker, high performing, undergraduate and graduate students in the British Academic Written English corpus-BAWE (Nesi et al., 2008) with a reference corpus of non-academic student writing. They identified 591 “academic” words (427 lemmas) used productively in the entire corpus of BAWE, which was six times smaller than the AVL itself, suggesting that the number of words students who receive exemplary scores use for writing can be considerably smaller than the number they access for reading and may differ from the range of academic words used by professional researchers, as in COCA. An implication is finding that the memorizing of extensive lists such as the AVL might be unnecessary for effective academic writing at the student level.

Nevertheless, while EAP teachers and students in pre-academic programs may welcome smaller vocabulary wordlists for pedagogical purposes, in the case of graduate level students, such limited lists may be insufficient, as the students also need to be able to read widely and at times deeply in the literature of their field. A gap between receptive and productive academic vocabulary can negatively affect students’ academic literacy (Evans & Green, 2007). If a reader cannot readily recognize at least 95% of the words in a text, comprehension is compromised (Nation, 2013). Since writing at the graduate level is highly dependent on wide and in-depth reading of peer-reviewed articles and academic books, a student with limited vocabulary might have diffi-
culty understanding theory and research findings to apply to their own thesis writing.

Regarding productive vocabulary, graduate students writing theses are in a different position from students taking an exam or timed tests like IELTS or TOEFL, as thesis writers can refer to specialized vocabulary found in textbooks or journal articles, which relieves them of having to rely heavily on memory. On the other hand, having access to material to summarize may cause students to rely rather too much on the vocabulary in the sources. Keck (2006) noted that in paraphrasing, knowledge of synonyms and clausal structure was essential, and that L2 students tended to stay closer to the wordings in the source texts than L1 students. For ELLs, collocation and the choice of appropriate synonyms, antonyms, hyponyms, and hypernyms can present a challenge. Using vocabulary appropriately also involves control over inflectional and derivational forms of words. However, these lexical variations are not evenly distributed across academic fields or genres and knowing which ones are appropriate cannot always be intuited (Hyland & Tse, 2007). As well as using formal “academic” words and formulaic phrases that are common across many disciplines, student writers need to choose the most accurate technical vocabulary while avoiding vocabulary that may be considered too conversational, colloquial, or intimate (Hyland & Tse, 2007; Leedham & Cai, 2013).

Apart from attempts to identify the words students need to know for academic writing success, a few other studies have focused on what students can do with the vocabulary they have. The BAWE corpus (Nesi et al., 2008) has proved to be very fruitful for the exploration of NS and NNS use of vocabulary in academic writing. Nesi and Moreton (2013) found limited use of inanimate abstract “shell” nouns like aspect, category, change, difficulty, fact, method, problem, process, reason, and system by EFL/ESL writers. Leedham and Cai (2013) used the BAWE to investigate the use of linking adverbials such as besides and on the other hand by Chinese international students in the UK and found overreliance on a limited set of these discourse markers, possibly due to the influence of EAP textbooks previously used in China.

Compared with research on international students based in English speaking countries, empirical studies on the vocabulary of non-native speaker graduate students studying and writing in
English in an EFL environment are scarce. Annelie and Erman (2012) compared lexical bundles used in academic writing by British undergraduate NS with NNS of English studying linguistics in Sweden. They found the NS used a greater variety of 4-word bundles that overall contribute to idiomaticity, fluency and pragmatic appropriateness. Vuković Stamatović et al., (2020) compared the lexical richness of theses by Montenegrn graduate students in philology with those of NS graduate students in the US. The NNs used considerably more NGSL words and 32% fewer from the NAWL list than the NS, indicating more reliance on general-purpose words and less coverage of vocabulary in their field.

There is even less opportunity for EFL students to interact with English speakers outside class in Asian countries compared with countries like Sweden, so they have few chances to compare their performance with that of other speakers of English and this may lead to (a sometimes misplaced) lack of confidence in their own proficiency. To discover whether their concerns are justified is one of the motivations for this study. Specifically, we investigated the following questions:

1. What is the academic vocabulary profile in a corpus of non-native speaker graduate student theses in an EFL environment?
2. How does that profile compare with reference corpora of proficient native and non-native speaker academic writing?
3. What changes in the academic vocabulary profile of the thesis writers occur over the period of drafting?
4. How does students’ actual performance on vocabulary compare with their self-evaluation of the weaknesses and deficits in their academic writing skills?

Research methods

As previously stated, the impetus for the investigation of productive vocabulary in students from four cohorts (C1-4) came partially from entry surveys starting with Cohort 3. In this section we describe the methods we used for compiling corpora, choice of wordlists and concordancing software, and the methods we used for the investigation of productive vocabulary.
Compiling the MAL TESOL C1-4 corpora
After graduation, all students in the MAL TESOL program (n = 62) freely agreed to have their thesis drafts analyzed for research purposes. Previously graded early and final drafts of the introduction, discussion and conclusion sections of the MAL VN students’ theses were anonymized, cleaned of numerical material in tables and figures, and converted to .txt files. As some students wrote many drafts throughout their candidature, while others only submitted a “first” draft in the weeks before the final draft was submitted, the decision on which drafts to consider as “first” was based on the first complete draft submitted to the U.S. partner’s LMS during Thesis Seminar 4: Thesis Completion. Prior to this final seminar, most students had only submitted draft literature reviews and methods sections, and these were not included in the corpus. Literature reviews were excluded as they were likely to reflect students’ reliance on close paraphrase of source material rather than the students’ own ability to construct original arguments. Methods and Findings sections were excluded as these sub-genres were not in the BAWE linguistics graduate reference sub-corpora.

The resulting two corpora (C1-4 first and C1-4 final) consisted of a compilation of first and final introduction, discussion, and conclusion chapters from all the C1-4 students (n = 62). The C1-4 corpora and the two BAWE reference corpora were parsed using UAM CorpusTool version 3.3x (O’Donnell, 2021) and lemmatized using Wordsmith Tools 8 (Scott, 2020).

The C1-4 corpus consists of topics in applied linguistics, specifically TESOL. The choice of specialized thesis topics in C1-4 tends to reflect the students’ own concerns, with written or spoken vocabulary (n = 11) by far the most popular single topic for research. Other topics included: speaking (n = 6), use of first language in the classroom (n = 5) listening (n = 4), pronunciation (n = 4), pragmatics (n = 2), writing (n-7), grammar (n-3,) reading (n = 2) some combination of language skills (n = 12), or other topics such as students’ or teachers’ strategies, attitudes or motivations (n = 6).

Choice of reference corpora
It was difficult to find a truly comparable corpus of non-native graduate student writing in applied linguistics (TESOL) for reference. The
closest freely available corpus we could find was the BAWE corpus of highly rated undergraduate and graduate student writing in several disciplines (6,506,995 running words). The BAWE graduate linguistics sub-corpus consisting of level 4 graduate student papers is considerably smaller (33 papers; 88,865 running words). There are two data subsets: NS, speakers of English as L1 (14 papers, 38,222 running words; mean text length 2,730 words), and NNS, students whose L1 is other than English, living and studying in the UK where they are immersed in the target language (19 papers, 50,643 running words; mean text length 2,665 words.) The BAWE texts sub-corpus texts are more narrowly focused on formal linguistics than the C1-4 texts. The genres are predominantly essays and critiques, so there is little language related to research methods and analysis of data. To make for a closer comparison, our analysis of C1-4 focused only on the draft and final versions of the theses’ introductions, discussions, and conclusions. These chapters are most likely to contain students’ most diverse vocabulary, with less reliance on the vocabulary of source texts, or the more formulaic vocabulary found in the methods and findings chapters.

For comparison of BAWE NS and NNS and C1-4 student texts against professional published academic research writing, we used a subset of COCA Academic (COCAA) as a second reference corpus: COCAA totals 120 million words and was too large for our purposes. From COCAA we extracted COCAA 2010-2012, a smaller representative subset of 958,8174 running words/tokens and 231,206 types as the entire COCAA corpus was too large for our concordancing software to handle. The COCAA 2010-12 subcorpus consists of whole papers, mostly from academic journals from nine disciplines so it would be expected to contain a wider range of technical words than either the C1-4 texts or the BAWE texts and would provide a reference point for target professional writing that could not be obtained from the BAWE graduate student texts.

Choice of wordlists and concordancing software
Earlier wordlists such as AWL (Coxhead, 2000) are additional to the first 1,000 or 2,000 words in the language and we assumed graduate level students of EAP would know these words as they account for 70-90% of words encountered in reading general texts. Nation (2013, p. 30)
estimated that knowing the GSL plus the AWL would together account for 86.1% coverage of academic texts for receptive use.

The Contemporary American English (COCA\textsuperscript{3}) general wordlist is based on COCA (now up to one billion words), up-to-date, and based on a wide range of spoken and written genres including ~ 28% academic. The New Academic Vocabulary List (henceforth, AVL) from Gardner and Davies (2014) is derived from the COCA academic subcorpus (COCAA) and consists of the 20,000 most common lemmas in COCA. The 3,000 core academic lemmas in the AVL list occur at least 50% more frequently per million words in COCAA than in the overall COCA corpus. They also have a good “dispersion” across nine key academic domains with at least 20% of the “expected” frequency in at least seven of those domains and no more than three times as expected in any of the nine domains. This means that the core academic lemmas are the kinds of words students could be expected to meet and use frequently in a wide range of disciplines. Because of the dispersion and frequency rules for inclusion in AVL, many of these words are also part of the first 3,000 words in COCA. Unlike the AWL, the AVL is not built on top of a general service list of the 2,000 most common words in the language. The AVL has higher representation in COCAA, at 13.8% compared with the AWL at 7.2% (Gardner & Davies, 2014).

One limitation of COCAA is that none of the nine academic domains represent texts specific to linguistics. The closest fit was a combination of education, social sciences and humanities. Therefore, we also compared the BAWE and C1-4 corpora against the ALTL, a new lemmatized list of 332 words from a 10,781,188 running word corpus of well-known applied linguistics and English language teaching textbooks (Gholaminejad & Sarab, 2020). The ALTL wordlist has 7.1% coverage in this corpus. Although this new list had not yet been critically evaluated in the literature, it seemed to be a relevant discipline specific list for our MAL TESOL students as it covers core topics in the program such as general linguistics, teaching methods and materials, research methods, testing and assessment, English for Specific Purposes, and discourse analysis.

\textsuperscript{3} See www.english-corpora.org/coca
We used Laurence Anthony’s AntWordProfiler 1.5 (2020) to come up with a lexical frequency profile (LFP) (Laufer & Nation (1995), of productive vocabulary in the student writing corpora, we compared overall frequency of lemmas (both core vocabulary and technical vocabulary) listed in the new Academic Vocabulary List (AVL) (Gardner & Davies, 2014). The percentage coverage of the AVL core wordlist (3,000 lemmas) and the AVL discipline-specified wordlist (2,456 lemmas from education, humanities and social sciences that occur at least three times as frequently in those domains than in COCA-Academic). Standardized text-to-token (STTR) ratios were calculated to get a general idea of the density of these words in the corpora, and to compensate for distortion due to varying lengths of texts (Breeze, 2008).

Results

We examine the entry survey results first, and then the results of the corpus analyses.

Results of the entry survey

Previous Experience with Academic Writing in English

Students’ responses to the multiple-choice question, “What is the longest piece of academic writing you have ever completed?” are presented in Figure 1.
The length of the master’s thesis required by the program equates to 8,750-12,500 words. As Figure 1 makes clear, only 7 of the 38 students in the sample, or fewer than 20%, had ever written a piece of academic writing that comes close (4 students reported that their longest piece of writing was 5,000-10,000 words, and 3 students had written 10,000-20,000 words). Almost half the sample (17/38 students) had never written a piece of more than 3,000 words.

In another question, students were asked to evaluate their academic English skills (listening, speaking, reading, and writing as used for understanding lectures, giving presentations, reading textbooks and academic articles, and writing academic papers) on a Likert scale of very weak (1) to very strong (6). One-way ANOVA and a post-hoc Tukey HSD test were performed to compare students’ rankings of the four skills. The results are presented in Tables 1 and 2.
Table 1

Results of a one-way ANOVA comparing student ratings of the four academic language skills

<table>
<thead>
<tr>
<th>Skill</th>
<th>M</th>
<th>SD</th>
<th>N</th>
<th>p-value</th>
<th>F-statistic</th>
<th>df</th>
<th>Source: Own elaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening</td>
<td>4.68</td>
<td>0.81</td>
<td>38</td>
<td>&lt;.00001</td>
<td>10.56</td>
<td>Between-groups = 3</td>
<td></td>
</tr>
<tr>
<td>Speaking</td>
<td>4.55</td>
<td>0.86</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>4.47</td>
<td>0.80</td>
<td>38</td>
<td></td>
<td></td>
<td>Within-groups = 148</td>
<td></td>
</tr>
<tr>
<td>Writing</td>
<td>3.74</td>
<td>0.46</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As can be seen in Table 1, the mean ratings of listening, speaking and reading skills fall between 4 (somewhat strong) and 5 (strong). However, the mean rating of writing is between 3 (somewhat weak) and 4 (somewhat strong). The one-way ANOVA shows a statistically significant difference among the mean ratings. Tukey’s Honestly Significant Difference (HSD) pairwise comparisons (not shown), revealed statistically significant differences between the mean rating of writing and the mean ratings of the other three skills.

One of the open-ended questions in the survey asked, “What is your biggest challenge in academic writing?” As a result of thematic analysis, several categories emerged from the answers. As can be seen in Figure 2, while students identified several areas of difficulty, vocabulary stands out as mentioned the most often (the totals in Figure 2 exceed 38 because some students identified more than one area of difficulty).
The survey results confirm that the students identified a number of shortcomings in their scholarly writing which they characterized primarily as problems with vocabulary, especially what students referred to as “academic” vocabulary. This category included the following subcategories, illustrated with quotations from student responses:

**Lack of vocabulary:**
- “[M]y vocabulary range, especially formal or academic words” (C3).
- “I do not have much academic vocabulary to use for my writing, so my academic writing is quite simple” (C3).
- “My biggest obstacle in academic writing is lexical resources” (C4).

**Word choice, accuracy, or appropriateness:**
- “How to use accurate words” (C3).
- “It is how to use the word properly” (C4).
- “Word choices” (C4).

“Grammar and sentence structure” also included specific references to academic writing:
- “[S]tructures used in academic writing are also the challenges for me” (C3).
I don’t know how to write formally, for example, which structures [...] to use” (C4).

“[A]dvanced grammatical structures” (C4).

The vaguer notion of “language/expression” was also of concern. Some students understood that they lacked the ability to write “academically” without being able to identify any specific aspects of academic writing:

“Academic language has always been my biggest problem in academic writing” (C3).

“[H]ow to write academic essays appropriately and beautifully and especially suitable for the Graduate program” (C3).

“[A]cademic writing style” (C4).

Considering that the surveys were conducted prior to taking any courses in the MA program, students’ descriptions of their perceived linguistic deficiencies tended to be vague or uncertain. From the professors’ perspective, the early drafts of many of our students seemed repetitive, over-reliant on a “lexical teddy-bears” - favorite words and phrases (Hasselgren, 1994), and alternating between highly formal and more common “conversational” words and phrases. Similar concerns had been raised by Breeze (2008). In choosing to focus on the students’ productive vocabulary, we decided first to focus on breadth (number and variety of words) to obtain a baseline for comparison with other graduate student writing rather than depth, which takes in phonemic, morphological and syntactic constraints as well as deeper level word associations (Crossley et al., 2012). Additional analysis on lexical richness including depth will be a subject of further investigation.

Results of the corpus analysis

Questions 1 and 2 sought to establish a lexical frequency profile (LFP) of our students as a whole at the beginning of the thesis writing process compared with the reference corpora and Question 3 explored how much change in vocabulary use occurred over the period of drafting, which ranged from 6 months to one year.

LFP (Laufer & Nation (1995) measures the lexical richness of the student texts and the productive size of the learners’ vocabulary. In other words, we quantified the extent to which the students in C1-4
made use of the lexical resources of the language productively in their writing and compared the vocabulary of our students with that of the NS and NNS BAWE linguistics subcorpora. By comparing the percentage of words from the AVL used by both populations and comparing both against the vocabulary used by professional writers in the COCAA 2010-12 sub-corpus we looked for an indication of the breadth of our students’ vocabulary.

We first established a benchmark for comparison of COCAA professional published writers, the BAWE NS and NNS, and our C1-4 writers. Using WST 8 (Scott, 2020) we calculated the total number of words and tokens in each corpus. The type to token ratio (TTR) was standardized per 1,000 words to compensate for the fact that longer texts will have a greater tendency for words to recur (Breeze, 2008; Coxhead et al., 2010). This is expressed as a percentage in Table 2 below.

Table 2
Overall results for all student corpora compared with COCAA (WST8)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Total types</th>
<th>Total tokens</th>
<th>STTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCAA 2010-12</td>
<td>144,444</td>
<td>9,588,174</td>
<td>41.4%</td>
</tr>
<tr>
<td>BAWE NS</td>
<td>4,565</td>
<td>38,223</td>
<td>39.87%</td>
</tr>
<tr>
<td>BAWE NNS</td>
<td>5,181</td>
<td>50,200</td>
<td>39.07%</td>
</tr>
<tr>
<td>C1-4 first</td>
<td>8,186</td>
<td>207,961</td>
<td>36.65%</td>
</tr>
<tr>
<td>C1-4 final</td>
<td>8,488</td>
<td>245,111</td>
<td>36.09%</td>
</tr>
</tbody>
</table>

Source: Own elaboration

Table 2 provides a general overview of the five corpora in the study. As COCAA 2010-12 is a large and diverse corpus of mostly academic peer-reviewed journal articles, the overall Standardized TTR is high, probably due to the diversity of subject specific lexis. The BAWE NS and NNS are much smaller corpora, with many fewer types and tokens overall, yet the STTR is quite high at over 39%. This might suggest a narrower set of field-specific words, but relatively little repetition. By contrast, the C1-4 corpora were larger than BAWE, covering different topics in applied linguistics, with a more diverse stock of discipline specific types. The slightly lower STTR of around 36% suggested there might be a slightly higher degree of repetition. However, it is important to note that at best, TTR is a rough measure of lexical density, and says
very little about the quality of the writing or the appropriacy of the vocabulary (Scott, 2020).

There was no statistically significant difference in mean STTR between the first and final drafts of our students’ theses \((t = -1.267894, p = .22552)\). Comparing the mean STTR of the C1-C4 final drafts with the mean STTR of BAWE NS and BAWE NNS corpora yielded no statistically significant differences, either: \((F = 0.90108; p = 413331)\). In other words, when it comes to STTR as a measure of lexical richness, the texts in C1-4 do not differ significantly from the BAWE corpora, which also do not differ much from each other or COCAA. Nation (2013) argued that by itself the number of different words a writer uses does not provide the whole picture and a more detailed analysis was needed to fully understand why so many of our students suspected their vocabulary was inadequate for thesis writing.

We analyzed the C1-4 against the reference corpora using AntWordProfiler (AWP) to discover the percentage coverage of the AVL core academic wordlist (3,000 lemmas) and the first 2,500 words of the AVL discipline specific wordlist for education, social sciences, and humanities. When analyzing using wordlists that overlap, as is the case here, AWP only includes in the second level those words which were not included in the first, so some core words that are common in Education are already counted in the core academic list.

<table>
<thead>
<tr>
<th>Wordlist</th>
<th>Types</th>
<th>Type%</th>
<th>Tokens</th>
<th>Token%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVL-core-academic</td>
<td>2,763</td>
<td>1.91</td>
<td>1,196,978</td>
<td>12.48</td>
</tr>
<tr>
<td>AVL-Ed, Soc. Sci, Hum-Specific</td>
<td>2,238</td>
<td>1.55</td>
<td>1,340,816</td>
<td>13.98</td>
</tr>
<tr>
<td>Other</td>
<td>139,443</td>
<td>96.54</td>
<td>7,050,380</td>
<td>73.53</td>
</tr>
<tr>
<td>Total</td>
<td>144,444</td>
<td>100.00</td>
<td>9,588,174</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Source: Own elaboration

Table 3 shows that in the COCAA 2010-12 reference corpus, coverage of AVL core academic and technical/domain (education, social sciences, and humanities) specific lists accounted for 26.46% of all tokens, which is close enough to the overall percentage (28%) for the entire COCAA subcorpus to make it a reasonable reference subcorpus.
for our purposes. Unsurprisingly, given the range of disciplines and research topics, 73.5% of tokens and 96.5% of types were words (Other) not found in the AVL wordlists.

We next compared the number and percentage frequency of (lemmatized) word types and tokens among the BAWE and C1-4 student corpora with COCAA to get an idea of the overall productive vocabulary students used, specifically what percentage of the vocabulary (by type and token) in each student corpus was from the first 3,000 “core” academic AVL wordlist, and what percentage was from the AVL education, social sciences and humanities list.

Table 4

<table>
<thead>
<tr>
<th>Wordlist</th>
<th>COCAA 2010-12</th>
<th>BAWE NS</th>
<th>BAWE NNS</th>
<th>C1-4 first drafts</th>
<th>C1-4 final drafts</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVL-core-academic</td>
<td>12.48</td>
<td>15.88</td>
<td>15.13</td>
<td>14.45</td>
<td>14.3</td>
</tr>
<tr>
<td>Other</td>
<td>73.53</td>
<td>69.61</td>
<td>70.06</td>
<td>67.8</td>
<td>67.57</td>
</tr>
</tbody>
</table>

Source: Own elaboration

Table 4 shows BAWE NS and NNS had a slightly higher percentage of core academic tokens while C1-4 had a noticeably higher percentage of additional discipline specific lexis. As explained earlier, longer texts in C1-4 might have contributed to this difference.

Table 5

<table>
<thead>
<tr>
<th>Wordlist</th>
<th>COCAA 2010-12</th>
<th>BAWE NS</th>
<th>BAWE NNS</th>
<th>C1-4 first drafts</th>
<th>C1-4 final drafts</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVL-core-academic</td>
<td>1.91</td>
<td>22.69</td>
<td>19.44</td>
<td>16.39</td>
<td>16.11</td>
</tr>
<tr>
<td>AVL-Ed, Soc.Sci, Hum-Specific</td>
<td>1.55</td>
<td>9.22</td>
<td>9.5</td>
<td>7.67</td>
<td>7.53</td>
</tr>
<tr>
<td>Other</td>
<td>96.54</td>
<td>68.08</td>
<td>71.07</td>
<td>75.93</td>
<td>77.11</td>
</tr>
</tbody>
</table>

Source: Own elaboration
When considering AVL core academic types, BAWE (especially NS) had a higher percentage in both AVL categories, but especially in the academic core. C1-4 had a noticeably higher percentage of “other” words, especially in final drafts. A chi-square test confirmed a statistically significant difference between BAWE NNS and C1-4, with C1-4 using a slightly lower percentage of AVL and discipline specific tokens and a higher percentage of “other” words, but the effect size was small. “Other” could have been either from the most common general words in COCA or (less likely) “rarer” words in the lexicon. To check this, we compared the percentage of vocabulary student writers used from the first 2,000 words of COCA, which include function words like of and everyday words like language, English, learning and teaching. Both NS and NNS in BAWE used around 63% of tokens and 23% of types; by comparison, C1-4 used 64% of tokens but only 15% of types. This shows that C1-4 drew on a much narrower set of basic vocabulary and repeated those words more.

Question 3 asked what changes had occurred between drafts. C1-4 first drafts (n = 62) totaled 207,961 running words (tokens) with a mean length 3,354 words (tokens). C1-4 final drafts (n = 62) 245,111 running words (tokens); mean length 3,953 words (tokens). Between the first and final drafts students increased overall token count by approximately 17.34%. There was considerable variation among students in every cohort between their first and second drafts. Table 6 shows results for Cohort 1 as an example.

Table 6  
Variation in number of types and tokens in Cohort 1

<table>
<thead>
<tr>
<th>Draft</th>
<th>Mean no. of types</th>
<th>St. Dev. Type range</th>
<th>Mean no. of tokens</th>
<th>St. Dev. Token range</th>
</tr>
</thead>
<tbody>
<tr>
<td>First draft intro</td>
<td>382.2</td>
<td>170.4</td>
<td>160-896</td>
<td>1050.9</td>
</tr>
<tr>
<td>Final draft intro</td>
<td>411.8</td>
<td>102</td>
<td>229-608</td>
<td>1158.8</td>
</tr>
</tbody>
</table>

Source: Own elaboration

Table 6 shows Cohort 1 first draft introductions ranged from 284-3743 tokens, and 160-896 types, while for final draft introductions, the range was 481 to 2385 tokens, and 229 to 481 types. Some students
shortened their final draft introductions considerably, shifting some material into other chapters such as the literature review.

Overall, the length of C1-4 draft introduction, discussion and conclusion increased 17.34% in terms of tokens, but not much in terms of type; the students collectively added a total of only 302 “academic” words, including 25 from AVL core and 11 from AVL discipline specific. This indicates that they did not re-word their texts very much between drafts. A lower percentage of AVL core words compared with BAWE again suggests they relied on a relatively small pool of familiar vocabulary and may have repeated words more than BAWE students.

There is no AVL discipline specific wordlist that refers only to linguistics or applied linguistics. The closest match is in education, social sciences, and humanities (2,456 lemmas). We therefore also compared the coverage in the student corpora of the ALTL which has 332 lemmas covering 7.1% of text in applied linguistics and English language teaching textbooks (Gholaminejad & Sarab, 2020). The results were as follows:

Table 7
BAWE NS Lexical profile statistics ALTL

<table>
<thead>
<tr>
<th>Wordlist</th>
<th>Types</th>
<th>Type%</th>
<th>Tokens</th>
<th>Token %</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALTL wordlist</td>
<td>239</td>
<td>5.24</td>
<td>1904</td>
<td>4.98</td>
</tr>
<tr>
<td>Other</td>
<td>4,326</td>
<td>94.76</td>
<td>36,319</td>
<td>95.02</td>
</tr>
<tr>
<td>Total</td>
<td>4,565</td>
<td>100.00</td>
<td>38,223</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Source: Own elaboration

Table 8
BAWE NNS Lexical profile statistics ALTL

<table>
<thead>
<tr>
<th>Wordlist</th>
<th>Types</th>
<th>Type%</th>
<th>Tokens</th>
<th>Token%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALTL wordlist</td>
<td>264</td>
<td>5.1</td>
<td>2,463</td>
<td>4.91</td>
</tr>
<tr>
<td>Other</td>
<td>4,917</td>
<td>94.9</td>
<td>47,737</td>
<td>95.09</td>
</tr>
<tr>
<td>Total</td>
<td>5,181</td>
<td>100.00</td>
<td>50,200</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Source: Own elaboration

Tables 7 and 8 show that in the BAWE, NS students used almost the same percentage of words from the ALTL list (5.24%) as did NNs (5.1%).
Tables 9 and 10 show that for C1-4, the percentage of words from the ALTL list barely changed between first and final drafts. Coverage of the LTAL academic vocabulary list was approximately 4.5% of all tokens in first and final drafts, with a range of 294-298 types, and was close to that of the BAWE students with 4.9% of all tokens and 239-264 types. This indicates that C1-4 coverage of the 332 academic lemmas commonly found in linguistics textbooks was comparable.

Discussion and conclusions
This preliminary study sought to discover an empirical baseline for investigation of the vocabulary level of students in an MA Linguistics TESOL program in an EFL environment. Quantitively, the productive vocabulary of these non-native academic thesis writers in applied linguistics compared favorably with both NS and NNS graduate linguistics students in the BAWE corpus, especially in the use of technical lexis of their field, though our students seemed to be working from a more limited set of vocabulary both from the first 2,000 words in COCA as well as core academic lexis in the AVL. Questions 1, 2, and 3 were partially informed by the lack of published studies of the productive vocabulary of graduate student writers in an EFL environment, while Question 4 sought to address issues of vocabulary raised in the MAL (TESOL) entry survey.
Regarding Questions 1 and 2, the corpus analysis suggests our MAL TESOL students were quite familiar with academic lexis and the basic terminology of their discipline. The very similar coverage of the ALTL between BAWE (who were all considered exemplary student writers) suggests that the breadth of productive vocabulary of these non-native academic thesis writers in an EFL context compared favorably with both NS and NNS graduate linguistics students in the (ESL) BAWE corpus. With respect to both the first 2,000 words in COCA and core AVL coverage, C1-4 used a smaller range but repeated them more often. This finding aligns with previous studies that report lower lexical variety and greater repetitiveness in L2 compared with L1 writers (Crossley et al., 2012; Staples & Reppen 2016; Yu, 2009) and further suggests that the writing of these international students has yet to attain the lexical richness found in professional writings by seasoned academics or even native speaker graduate students. However, this expectation for non-native speaker students may be unreasonable (Malmström et al., 2018).

Regarding Question 3, while there was 17.34% expansion in the overall length of writing between first and final drafts there was negligible expansion in the number of word types. These results were not unexpected, as our classroom experience was that most students did little more than superficial editing at the vocabulary and syntactic level and relied more on trying to correct surface errors pointed out by the instructor. The reasons for this may be related to the short time period for new vocabulary learning to take place, as well as students’ limited confidence in using inflectional and derivational forms or their unfamiliarity with register-appropriate synonyms, antonyms and collocates.

Turning to Question 4, we found that C1-4 students’ productive use of vocabulary in terms of breadth compared favorably with the reference corpora and was better than their self-evaluation of their weaknesses and deficits suggested; they had not only identified vocabulary as their top concern in preparing a thesis but had also chosen vocabulary as by far the most popular single topic for research, accounting for 11 of the 62 theses. Although the original survey did not explicitly explore students’ understanding of the meaning of “vocabulary,” we gained an idea of what students understood by vocabulary by referring to qualitative responses they made in the survey and by examining the definitions of vocabulary some wrote in thesis chapters which were included.
in this study. Some students failed to adequately define vocabulary, while others did show a more sophisticated grasp of the meaning of vocabulary based on their reference to the literature. While not necessarily defining vocabulary, most students referred to aspects such as “semantic, morphological, pragmatic, and syntactic knowledge.” The qualitative comments made in the survey point to issues that extend beyond mere knowledge of words that appear on the AVL or ALTL wordlists and highlight the fact that depth of vocabulary and syntax are inextricably related in academic writing.

In conclusion, our preliminary study demonstrates that these graduate students in an EFL context compared very favorably with the BAWE corpus of exemplary NS and NNS graduate students on breadth of vocabulary, and better than Montenegrin graduate students reported by Vuković Stamatović et al., (2020). Our lexical profiling suggests that the C1-4 students have adequate coverage of technical terms in their field as well as core academic words represented in the AVL and ALTL wordlists, however their writing tends to be repetitive, especially of the most common words in English. That said, the extent to which studying either core or specialized academic wordlists can help non-native graduate students produce more sophisticated and native-like academic writing remains to be seen. Most research so far has been on receptive rather than productive use of these lists (Breeze, 2008) and while an extensive receptive vocabulary is undoubtedly an advantage in graduate-level studies, some recent work suggests that students can still manage to write effectively with a smaller productive vocabulary (Malmström et al., 2018).

We recognize several limitations of this study. First, disciplines differ extensively in the vocabulary they favor (Durrant, 2016; Hyland & Tse, 2007). Given the differences between formal linguistics and applied linguistics, the BAWE reference corpus may not provide the best comparison with the C1-4 corpus, yet it is the most readily available free corpus of graduate student writing with any relation to linguistics. Second, while AVL seems to be the best wordlist for now, being current and based on a subset of the billion-word COCA corpus, continuing interest in specialized corpora will no doubt lead to more precise wordlists. For example, the ALTL seems promising but has not been widely tested. Third, space limitations for this paper have precluded
analysis of the depth of vocabulary usage by our graduate students. Another paper will examine whether students’ concerns about accuracy or appropriateness of word choice is justified and will consider use (or “misuse”) of register-inappropriate words and phrases, discourse markers, synonyms and collocations, unnecessary repetition, selection of grammatically inappropriate derivational suffixes, and over-reliance on “lexical teddy bears” (Hasselgren, 1994).

The number of students in the study (n = 62) does not allow for generalization to other populations of graduate student writers, but the results provide a launching point for deeper investigations into the nature of vocabulary use by international graduate students studying in an EFL environment.

References


The student reference corpus data in this study come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics, Warwick), Paul Thompson (formerly of the Department of Applied Linguistics, Reading) and Paul Wickens (School of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800). www.coventry.ac.uk/bawe/