Milica Vuković Stamatović,

Vesna Bratić, Igor Lakić

University of Montenegro, Montenegro

2020, Vol. 17 (2), 101-113(284)

revije.ff.uni-lj.si/elope

<https://doi.org/10.4312/elope.17.2.101-113>

UDC: [808.1:001.8]:378(497.16:73)

Vocabulary of L1 and L2 Graduation Theses Written by English Philology Students: Academic Writing of Montenegrin and US Students Compared

ABSTRACT

The paper explores the lexical profile of graduation theses written by the students at the University of Montenegro and compares it against that of BA theses authored by native speakers of American English. We study their lexical level (LFP method), lexical variation (sTTR method), and share of academic vocabulary according to the New Academic Word List. We aim to determine how different L2 academic writing is and where the lexical differences may lie, so that pedagogical recommendations can be made. The results show that the Montenegrin theses are readable at 4,000 words, which means that CEFR B2 learners can read them at a reasonable level. In contrast, the theses written by native speakers can be read at 7,000 words, i.e. only by those commanding good C levels. As this is in line with our expectations, we conclude that the Montenegrin theses display a sufficient vocabulary size. Since the students still underuse academic vocabulary, we recommend that more emphasis should be placed on it during their studies.

Keywords: lexical profile, lexical variation, LFP, sTTR, academic vocabulary, NAWL

Besedišče prvega in drugega jezika v diplomskih nalogah študentov angleškega jezika: Primerjava akademskega pisanja črnogorskih in ameriških študentov

POVZETEK

Članek obravnava leksikalni profil diplomskih nalog študentov Univerze v Črni gori in ga primerja s profilom nalog govorcev ameriške angleščine. Razišče njihovo leksikalno raven (metoda LFP), leksikalno variacijo (metoda sTTR) in delež akademskega besedišča po seznamu New Academic Word List. Ugotavljamo, kako drugačno je akademsko pisanje v drugem jeziku in kje se lahko pojavljajo leksikalne razlike. Iz ugotovitev izhajajo priporočila za pedagoško obravnavo. Rezultati kažejo, da so črnogorske naloge berljive na ravni 4000 besed, kar pomeni, da jih lahko razmeroma dobro berejo učenci na ravni B2 po SEJO. Po drugi strani pa naloge, ki so jih napisali rojeni govorniki, zahtevajo poznavanje 7000 besed, ki je dosegljivo le učencem na ravneh C. Rezultat je pričakovan, zato zaključimo, da črnogorske diplomske naloge kažejo zadostno velikost besedišča. Ker črnogorski študenti premalo uporabljajo akademsko besedišče, priporočimo večji poudarek na njegovi obravnavi tekom študija.

Ključne besede: leksikalni profil, leksikalna variacija, LFP, sTTR, akademsko besedišče, NAWL

1 Introduction

In this paper, we study the lexical profile of the diploma papers written by the students graduating from the Department of English Language and Literature, at the University of Montenegro. The aim is to explore their writing in terms of its lexical level and variation, as well as their command of academic vocabulary, and observe how it compares against the academic writing of the native students graduating from the same field. Therefore, this study is both descriptive and contrastive in nature and, based on the results, we aim to draw some pedagogical implications related to developing vocabulary and writing skills in these students.

To achieve these aims, we use a corpus linguistic methodology. Namely, to measure the vocabulary level or load, we employ the Lexical Frequency Profiling (LFP) method, which will also be used to measure the coverage of academic vocabulary in our corpus. To calculate lexical variation, we use the standardised Type-to-Token Ratio (sTTR) method. The resulting lexical profile of our corpus is compared against the lexical profile of a corpus of bachelor theses written by native speakers of American English to determine where the lexical differences may lie.

2 Theoretical Background

In this section of the paper, we briefly cover two broad topics – the corpus linguistic methods and how L2 writing generally compares against L1 writing in terms of vocabulary.

2.1 Corpus Linguistics

It was not before the early 1980s that the name of the discipline dealing with the creation and description of corpora – *corpus linguistics* – entered general use (McCarthy and O’Keeffe 2010, 5). Corpus linguistics falls within the umbrella concept of *descriptive linguistics*, which can yield objective, quantitative results, the analysis of which typically continues using other linguistic methods, so as to interpret the obtained data in a qualitative way. Therefore, corpus linguistics by no means represents an end in itself: it is a research tool which can be used in most linguistic disciplines based on exploring authentic language, i.e. language in use.

The basic procedure in corpus linguistics involves work with sets of authentic texts called *corpora*, the largest of which amount to as many as half a billion words. These can be processed automatically via various software, producing vast amounts of various data on language patterns and behaviour. The next research step is to statistically interpret the data thus obtained, making use of more or less complex techniques in the process. Patterns are sought and meaning(s) are interpreted as a function of context. There are several methodologies in corpus linguistics, which prove most useful when examining vocabulary and morphological categories. This method boasts the advantages of objectivity and systematicity in processing vast amounts of authentic data.

In the following subsections we briefly present the corpus linguistic methods which will be employed in this study, as stated in the introduction.

2.1.1 Lexical Frequency Profiling and Word Lists

The method of Lexical Frequency Profiling (LFP) was developed in 1995 by Laufer and Nation, and is used for quantifying the lexical richness of a text as well as the productive size of the learner's vocabulary. The method is based on the following procedure: a corpus is "entered" into a specialised program alongside one or several word lists (e.g. these lists can feature high-frequency vocabulary, academic vocabulary, technical vocabulary, etc.). The program calculates the amount of coverage of each of these lists in the corpus loaded. The obtained results can then be compared to those for other corpora, which reveals how lexically complex a certain corpus is in comparison to others.

Being the best-known frequency-based measure of vocabulary, the LFP is very often used in EFL research and teaching, most typically to determine the lexical complexity of certain texts. Although it is not the sole method for calculating lexical richness, the LFP has been found to produce results that to a great extent match those obtained using other methods (Lindqvist Gudmundson, and Bardel 2013). Even though the results obtained in this way are numerical, which certainly contributes to their clarity and verifiability, the method has met with some criticism. The most prominent fault found is that it shows a bias towards receptive knowledge. In addition, when the lexical profile is "reduced" to just word frequencies, some "information loss" seems to be inevitable (Crossley, Cobb, and McNamara 2013). Nevertheless, over the course of the past two decades this method has been widely used (Cobb and Horst 1999; Morris and Cobb 2004; Read and Nation 2006; Douglas 2015 etc.).

Today word lists are normally compiled from large authentic corpora. The ones containing the most frequent general-purpose vocabulary are most typically employed in teaching General English. Others are more narrowly specialised for certain domains, such as academic word lists for tertiary-education cycle students or ESP learners' lists. They are typically employed as teaching and learning resources (Khani and Tazik 2013), as well as guidelines for developing EFL and ESP textbooks, curricula and courses (Wang, Liang and Ge 2008).

One word list in particular, the General Service List (GSL, West 1953), was influential for decades. It was extracted from a five million-word corpus manually and contains the most frequent 2,000 word families¹ of English. The GSL was widely used until quite recently, its updated replacements having been introduced as late as in 2013. The updated GSLs are known as the NGSLs, standing for the New General Service Lists (Brezina and Gablasova 2013; Browne, Culligan, and Phillips n.d.b) and they outperform the old GSL to a certain degree, as the original GSL is somewhat outdated now.

The Academic Word List (Coxhead 2000) was developed by using the GSL in such a way that the words represented in this word list were excluded, whereby an academic corpus of 3.5 million words was used as a basis for extracting the word list. It was found to cover about 10% of the words in most academic corpora with its 570 word families.

Just as the AWL was built on top of the original GSL, the NAWL (New Academic Word List; Browne, Culligan and Phillips n.d.a) was developed on top of the NGSL (Browne,

¹ A headword and all its inflected and derived forms, for instance, *know*, *knows*, *knowing*, *knew*, *known*, *knowledge*, *knowledges*, *knowledgeable*, *unknown*.

Culligan and Phillips n.d.b). Therefore, these can be used together, as is the case with the previous set of word lists (the GSL and the AWL). The NGSL was developed based on a 273 million-word subsection of the Cambridge English Corpus, whereas the NAWL was derived from a 288 million-word academic corpus. In the latter, the NGSL covers 86%, whereas the NAWL covers an additional 6%, according to the dedicated website of Browne, Culligan and Phillips (n.d.b).² As can be seen, the newer lists are based on much bigger corpora than the word lists produced earlier and should, therefore, provide more coverage in various texts.

In 2012, the enormous corpus which combined the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) gave birth to a set of frequency-based word lists (Nation 2020): 25 now exist, each featuring 1,000 word families. This set comes alongside four additional word lists, containing proper names, marginal words, transparent (non-hyphenated) compounds and abbreviations. Proper names are generally considered easy to recognise, and their coverage is often calculated together with the word lists containing content words to determine how readable texts are. Some authors also assume that abbreviations (which are typically explained in the texts in which they are used) and non-hyphenated compounds (which are generally easy to decipher from the meanings of their parts when found in context) do not add to the vocabulary load either (Hsu 2014; Vuković Stamatović 2019). The same might be said of marginal words in our case, given that they contain the letters of the alphabet, swear words and exclamations – in our corpus, there are neither swear words nor exclamations, and so the supplementary word list containing marginal words covers only the letters of the alphabet (for instance, “*the example D*”), and these can also be treated as non-words or words which do not add to the vocabulary load. Nation’s (2017) set of the described word lists can be used for determining how lexically demanding a corpus is.

2.1.2 TTR and sTTR

The most frequently used methodology for measuring lexical variation is that of the TTR, standing for Type-to-Token Ratio. *Tokens* represent the number of individual words used in a text (“running words”) and *types* refer to the number of the unique word forms in that text. The formula is sensitive to the size of the corpus (Kubát and Milička 2013) and, therefore, not very reliable when comparing corpora of different sizes. This has attracted different attempts to improve it. In his *WordSmith Tools*, Scott (2004) integrated a measure called standardised TTR (sTTR), in which a text is divided into chunks of equal length and the TTR is calculated for each of these chunks separately; the final sTTR result represents the average of all the ratios calculated for the equally-sized chunks.

TTR-based methods measure diversity, i.e. variability of the vocabulary. Nevertheless, variation cannot provide an insight into the level of the vocabulary, i.e. what type of words are used and how frequent these words are in general language. For example, a text may display substantial lexical variation and still comprise mainly mid-frequency general-purpose words. Having this in mind, we opted to use both the LFP method and the sTTR measure, as suggested above.

² See <http://www.newgeneralservicelist.org/nawl-new-academic-word-list>.

2.2 Vocabulary and L1 and L2 Writing

L2 writing has been found to vary significantly from L1 writing (Troia 2007). In terms of vocabulary, L2 writing was found to feature a lower lexical variation (Linnarud 1986), use fewer words (Staples and Reppen 2016) and contain errors arising from an inadequate selection of words (Sonomura 1996; Eckstein and Ferris 2018).

Nation (2013) finds that there are approximately 70,000 word families in English, and that an average high-school graduate will know about 20,000 of them. As for L2, Nation and Waring (1997) find that an adult ESL/EFL speaker typically knows fewer than 5,000 word families of English, even though he/she has learned this language for years. As can be seen, there is a big gap between the vocabulary size of native and non-native speakers of English, and vocabulary is certainly one of the major obstacles that non-native speakers face.

Milton (2010, 226) and Capel (2012) attempt correlating vocabulary sizes with the CEFR³ levels, and determine roughly the following correspondences: A1 level – 1,000 word families; A2 level – 2,000 word families; B1 level – 3,000 word families; B2 level – 4,000 word families; C levels⁴ – 5,000+ word families.

These numbers correspond to a learner's receptive knowledge of vocabulary. The best non-native speakers, such as doctoral students studying in English as a medium of instruction, are likely to know the most frequent 9,000 word families of English, Nation estimates (2013, 26).

When it comes to academic vocabulary, learning this has been the focus of teaching English for Academic Purposes (EAP), which is a common course found in the undergraduate curricula pursued by L2 speakers. Moreover, Snow and Ucelli (2009) argue that academic vocabulary is also more challenging than other registers for native speakers, and that it should be explicitly taught to them too.

3 Data and Procedure

Two corpora are used in the paper. One of them is a corpus of 24 diploma papers recently written by the students graduating in English Philology from the Faculty of Philology of the University of Montenegro (after four years of studies). Twelve of these graduation theses belong to the field of linguistics and the other 12 to the field of literary studies. This corpus comprises 180,020 tokens.

The second corpus comprises bachelor theses written by native English speakers – 12 of them are from the discipline of English linguistics and they were taken from the repository of the Department of Linguistics of the Ohio State University (Ohio, USA), while the other 12 are from English literary studies, and these were taken from the repository of the Middlebury College (Vermont, USA). This corpus comprises 324,554 tokens.

³ Common European Framework of Reference for Languages (Council of Europe 2001).

⁴ Milton (2010, 226) and Capel (2015, 4) do not make a distinction between the C1 and the C2 levels in terms of vocabulary size.

We use a computer programme called AntWordProfiler 1.4.1 (Anthony 2014), a freeware tool⁵ developed by Laurence Anthony, which is widely used for determining vocabulary load and corpora complexity.

The research questions we aim to answer are the following:

RQ1. How lexically rich are Montenegrin graduation theses written by the students of English philology in comparison with BA theses written by their English philology peers from the USA?

RQ2. What is the level of lexical variation in the Montenegrin graduation theses written by the students of English philology in comparison to that of the bachelor theses written by their English philology peers from the USA?

RQ3. How much academic vocabulary do Montenegrin English philology students use in their graduation theses in comparison to their US peers?

The results are presented in the next section.

4 Results and Analysis

When comparing the two datasets, one first notices the difference in the number of words the Montenegrin and the US graduation theses feature. An average Montenegrin thesis contains 7,500 words, which corresponds to an average length of an academic article published in a philological journal. On the other hand, the US theses prove to be much longer, featuring around 13,500 words on average. Of course, these differences are a matter of conventions and a consequence arising from the guidelines given to the students, but they might point to how much importance is given to the graduation piece of writing in the two education systems.

4.1 Lexical Level of Graduation Theses

We first present the coverages reached by Nation's (2017) set of word lists in the two corpora (Table 1). We use the version of these word lists found on Anthony's website.⁶ We present the results for the first nine word lists, for convenience, as the remaining 16 word lists featured very little frequency in the corpora.

When analysing the lexical frequency profile in L2 texts, researchers usually first look at how much of the text is covered by the most frequent words – typically, the first 2,000 words (as we have seen, the knowledge of these words generally corresponds to the A2 level according to CEFR). Generally, the higher this percentage is, the simpler the overall vocabulary will be, bearing in mind that less room will be left for the more 'difficult', i.e. generally infrequent words. The first 2,000 words together take up 83.13% and 77.89% in the Montenegrin and the US corpora of graduation theses, respectively, which is a first indication that the graduation writing of the native English speakers will be of a significantly higher level, as expected.

As the writing in question is certainly more advanced than the A2 level, i.e., it goes substantially beyond the most frequent 2,000 words of English, more measures need to be considered, and usually readability of texts is determined as a next step.

⁵ See <https://www.laurenceanthony.net/software/antwordprofiler/>.

⁶ See <https://www.laurenceanthony.net/software/antwordprofiler/>.

TABLE 1. Lexical level of graduation theses.

BNC/COCA word lists	Montenegrin graduation theses		US graduation theses		Examples of frequent words from the word lists
	%	Cum. %	%	Cum. %	
Proper names	3.3	3.3	3.34	3.34	English, America, Ann, Midland, Faulkner, Scotland, Rosemary...
Marginal words	0.39	3.69	0.43	3.77	p, l, e, h, b...
Transparent compounds	0.17	3.86	0.2	3.79	southeast, forever, storyteller, afterthought, baseline, lifestyle, lifetime, notebook...
Abbreviations	0.13	3.99	0.2	4.17	al., tt, En., L1, www, http, UK, org, conj.
1,000 BNC/ COCA	73.96	77.95	67.88	72.05	the, be, of, and, to, read, word, when, can, use, write, at, other...
2,000 BNC/ COCA	9.17	87.12	10.01	82.06	language, speech, character, sentence, example, argue, introduce...
3,000 BNC/ COCA	6.78	93.9	7.51	89.57	text, discourse, novel, narrate, clause, distinct, notion, plot...
4,000 BNC/ COCA	1.38	95.28	2.17	91.74	utter, classification, linguistic, portray, prejudice, feminine, oppress...
5,000 BNC/ COCA	0.85	96.13	1.26	93	verb, cue, intelligible, patriarch, temporal, token, deviate...
6,000 BNC/ COCA	0.60	96.73	1.18	94.18	dialect, creak, syllable, conjunction, protagonist, heroine, accord, denote...
7,000 BNC/ COCA	0.64	97.37	0.82	95	vowel, dislocate, cohesion, mutilate, grammatical, psyche, trajectory...
8,000 BNC/ COCA	0.43	97.8	0.56	95.56	lexical, goblin, acoustic, pronoun, elliptical, chaste, permeable...
9,000 BNC/ COCA	0.22	98.02	0.36	95.92	slander, zombie, phonetic, stoke, entropy, atone, subjugate...

There are two figures in the literature taken as reading comprehension thresholds – the lower one is a 95% coverage (Laufer 1989), required for “reasonable” or “minimum” reading comprehension. So, how many words are needed to read the graduation theses from our corpus at the level of reasonable comprehension? The Montenegrin theses are read at 4,000 words, if one assumes an easy recognition of the words from the supplementary word lists, as argued in section 2.1.1 (these words are typically taken as not adding to the vocabulary load of a text). On the other hand, to read the US graduation theses at the level of reasonable comprehension, the first 7,000 words are needed, which points to much higher lexical demands. If this is correlated with the CEFR levels, it would mean that an EFL/ESL speaker who is at a B2 level could read the Montenegrin graduation theses, but that the C levels would be needed to read those written by the US students.

If we consider the results with regard to the vocabulary size of the Montenegrin graduating students, we can observe that 95% of their writing consists of the most frequent 4,000 words, and this is their minimum productive vocabulary size as displayed in writing (receptive vocabulary typically exceeds the productive one (Webb 2008)). In the remaining 5% of their writing they also use the more infrequent words, i.e. those from the higher frequency bands. Based on these two pieces of information, we can conclude that they, on average, have a vocabulary at the C levels, which is what is expected of them, given that they are graduating from English philology.

The ideal reading level is reached if one knows 98% of the words used in a text (Nation 2006). Such ideal reading is possible at 9,000 words when Montenegrin graduation theses are in question – which is within the reach of good non-native speakers. Nation (2013, 26) estimates that L2 students doing doctoral studies in English command a vocabulary of at least 9,000 words, and our results fit in with those estimates. However, for the US graduation theses, the whole of Nation’s set of word lists did not suffice. Beyond the first 25,000 words of English are typically some specialised words from various fields or some archaic ones. In this case, it would mean that a specialised vocabulary from the field of philology would be needed if one wanted to read the U.S. graduation theses at an ideal level of comprehension.

The differences between the two groups of writers are, of course, expected; however, our goal was to ascertain how big they exactly are. The Montenegrin graduation theses are written by non-native speakers at the C levels (most probably, C1), and can be read by others holding a similar level of English. With this level, the US graduation theses could be read with reasonable comprehension, but only an expert in the field, i.e. one with a good knowledge of the specialised philological vocabulary, could read them at an ideal level.

4.2 Academic Vocabulary in Graduation Theses

We aim to examine how much academic vocabulary Montenegrin students use in comparison to their peers who are native speakers of English. We use the NGSL and the NAWL for this purpose (Table 2), given that this is a recent set of word lists developed from an enormous corpus. These results are presented in Table 2 below.

TABLE 2. Coverage of NGSL and NAWL in graduation theses (%).

Word list	Montenegrin graduation theses	US graduation theses	Examples of the words from the word list
NGSL	85.81	80.23	the, be, of, and, to, reader, sentence, voice, different, know, how, text, come, same, show, listener, verb, look, here, although, say, first, such, through, among, develop, heart...
NAWL	2.33	3.43	classification, marker, non, publish, precede, temporal, aspect, portray, definite, render, resemble, conception, denote, thereby, progressive, commentary, subtle, sphere, repertoire...
Total	88.14	83.32	/

As can be seen in Table 2, there is more frequent general-purpose vocabulary in the writing of Montenegrin students. The differences range between five and six percentage points and are in line with the results from Table 1.

Browne, Culligan and Phillips (n.d.b)⁷ state that the NAWL covers 6% of the academic corpus from which it was derived. We find substantially less academic vocabulary in both corpora, probably resulting from the fact that students are novices at academic writing and not very experienced in it. However, an important finding is that there is considerably less academic vocabulary in the writing of Montenegrin graduating students – they used about 32% less of the NAWL vocabulary in comparison to the US students. And while the differences in the use of frequent and infrequent general vocabulary are expected, bearing in mind the fact that we are comparing native with non-native writers, the big differences in the use of academic vocabulary are less justified by this. Students studying philology are explicitly instructed in using this vocabulary throughout their education, and in the ESL literature it is generally advised to teach academic vocabulary to L2 tertiary students just after their having mastered the basic general vocabulary (for instance, the first 2,000 words of English) (Coxhead 2000). As the academic word lists represent the most frequent academic vocabulary, focusing on one of them as the learning target in undergraduate English philology studies would be the most efficient use of time. For instance, the NAWL list, with its 963 lemmas, is a feasible learning target to be aimed at in these studies, which typically last three to four years.

The gap in the amount of academic vocabulary between the two groups leads to the pedagogical recommendation that some more emphasis on academic vocabulary should be made in the teaching of Montenegrin students of English philology.

4.3 Lexical Variation in Graduation Theses

Finally, we look at how much lexical variation the writing of the two groups of students displays (Table 3).

⁷ See <http://www.newgeneralservicelist.org/nawl-new-academic-word-list>.

TABLE 3. Lexical variation in graduation theses (%).

Lexical variation measure	Montenegrin Graduation Theses	US Graduation Theses
sTTR for 10,000-Word Chunks	0.2	0.22

Given that the two datasets were of different sizes, we used the standardised TTR measure. As explained earlier, this ratio is obtained when the number of the unique word forms (types) is divided by the number of running words in a given segment (here we opted for 10,000-word-long chunks); the higher the ratio obtained, the more lexical variation a text displays. The results point to a slightly smaller lexical variation in the Montenegrin corpus, but the differences between the two groups are not that large (0.2 vs. 0.22). This finding points to how important it is to take different measures of lexical richness into account when attempting to compare different corpora – the TTR-based measure alone would not have pointed to a significant difference in the two datasets, although the researcher could assume that there would be significant differences. In this case, only the LFP measures pointed to the differences and confirmed the expectation that the writing of the L2 students would be less complex in terms of vocabulary.

5 Pedagogical Implications

The writing of Montenegrin students featured considerably less academic vocabulary in comparison to that of their US peers. Therefore, the Montenegrin students should be encouraged to master and use more academic vocabulary in their graduation theses, bearing in mind that they represent an academic genre governed by certain academic writing conventions, as well as a specific syntax. As with learning any type of vocabulary, providing students with sufficient exposure to it, i.e., input flooding, is essential. A problem here may be that undergraduate L2 students, in general, are not properly equipped to read authentic research papers in English until they have graduated, on account of their vocabularies, as well as their insufficient familiarity with advanced linguistics and literary studies. This problem may be overcome by exposing them to L1 undergraduate seminar and graduation papers, i.e., those written by their L1 peers – typically, these will be less complex in terms of content than a research article written by a scholar, but will expose them to authentic academic vocabulary and good models of academic writing.

We also recommend focusing on a specific academic vocabulary list – all the mentioned corpus-derived lists have been carefully produced to reflect the most frequent vocabulary with best range and distribution in various types of academic writing and across disciplines. They thus provide the largest coverage with the least number of words in academic writing, which makes focusing on them and teaching them strategically a very efficient use of time. They all contain between 500 and 1,000 words, and would thus be feasible learning targets as part of undergraduate L2 philology studies. Hirsh and Coxhead (2009) provide concrete suggestions of the methods of teaching specialised word lists effectively.

Some of the students will continue their postgraduate studies abroad or wish to publish in scholarly journals, and based on our results they would not be entirely ready for this. Therefore, the pedagogical aim should be to work in the direction of overcoming this shortcoming in their academic writing.

6 Conclusion

Based on the results presented above, we can say that the writing of the Montenegrin students displayed sufficient lexical richness, bearing in mind that they are non-native speakers. The amount of the vocabulary they used pointed to their having the C levels, according to the CEFR scale. Their vocabulary size was in line with the level of vocabulary that Nation (2013) suggests the L2 postgraduate students typically have. In addition, we discovered that the writing of Montenegrin students did not feature as much academic vocabulary as is expected in academic writing. Based on this, we recommend that more focus be placed on academic vocabulary and its use in academic writing in philological studies in Montenegro.

As the present study is based on a corpus of 24 Montenegrin graduation theses and the same number of US graduation theses, we note that a larger corpus would produce more reliable results. Bearing in mind that English philology is a diverse field, perhaps some differences in the linguistic and literary theses could be observed or these could be explored separately in further research. Related to this, we suggest a larger cross-comparison of graduation theses amongst additional groups of L1 and L2 students, which could also encompass theses from more universities. A larger-scale study could more reliably point to certain universal differences in the L1 and the L1 academic writing of the graduating students.

References

- Anthony, Laurence. 2014. *AntWordProfiler* (version 1.4.1). Tokyo, Japan: Waseda University.
- Brezina, Vaclav, and Dana Gablasova. 2013. "Is There a Core General Vocabulary? Introducing the New General Service List." *Applied Linguistics* 36 (1): 1–22. <https://doi.org/10.1093/applin%2fam018>.
- Browne, Charles, Brent Culligan, and Joseph Phillips. n.d.a. "The New Academic World List 1.0." New General Service List. <http://www.newgeneralservicelist.org/nawl-new-academic-word-list/>.
- . n.d.b. "The New General Service List." Accessed September 1, 2019. <http://www.newgeneralservicelist.org/>.
- Capel, Annette. 2012. "Completing the *English Vocabulary Profile*: C1 and C2 Vocabulary." *English Profile Journal* 3 (e1): 1–14. <https://doi.org/10.1017/S2041536212000013>.
- Cobb, Tom, and Marlise Horst. 1999. "Vocabulary Sizes of Some City University Students." *Journal of the Division of Language Studies of City University of Hong Kong* 1: 59–68.
- Council of Europe. 2001. *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- Coxhead, Averil. 2000. "A New Academic Word List." *TESOL* 34 (2): 213–38. <https://doi.org/10.2307/3587951>.
- Crossley, Scott A., Tom Cobb, and Danielle S. McNamara. 2013. "Comparing Count-Based and Band-Based Indices of Word Frequency: Implications for Active Vocabulary Research and Pedagogical Applications." *System* 41 (4): 965–81. <https://doi.org/10.1016/j.system.2013.08.002>.
- Douglas, Scott Roy. 2015. "The Relationship between Lexical Frequency Profiling Measures and Rater Judgements of Spoken and Written General English Language Proficiency on the CELPIP-General Test." *TESL Canada Journal* 32 (9): 43–64. <https://doi.org/10.18806/tesl.v32i0.1217>.
- Eckstein, Grant, and Dana Ferris. 2018. "Comparing L1 and L2 Texts and Writers in First-Year Composition." *TESOL Quarterly* 52 (1): 137–62. <https://doi.org/10.1002/tesq.376>.

- Hirsh, David, and Averil Coxhead. 2009. "Ten Ways of Focusing on Science-Specific Vocabulary in EAP Classrooms." *English Australia Journal* 25 (1): 5–16.
- Hsu, Wenhua. 2014. "Measuring the Vocabulary Load of Engineering Textbooks for EFL Undergraduates." *English for Specific Purposes* 33: 54–65. <https://doi.org/10.1016/j.esp.2013.07.001>.
- Khani, Reza, and Khalil Tazik. 2013. "Towards the Development of an Academic Word List for Applied Linguistics Research Articles." *RELJ Journal* 44 (2): 209–32. <https://doi.org/10.1177/%2f0033688213488432>.
- Kubát, Miroslav, and Jiří Milička. 2013. "Vocabulary Richness Measure in Genres." *Journal of Quantitative Linguistics* 20 (4): 339–49. <https://doi.org/10.1080/09296174.2013.830552>.
- Laufer, Batia. 1989. "What Percentage of Text Lexis Is Essential for Comprehension?" In *Special Language: From Humans Thinking to Thinking Machines*, edited by Christer Laurén and Marianne Nordman, 316–23. Clevedon: Multilingual Matters.
- Laufer, Batia, and Paul Nation. 1995. "Vocabulary Size and Use: Lexical Richness in L2 Written Production." *Applied Linguistics* 16 (3): 307–22. <https://doi.org/10.1093/applin/16.3.307>.
- Lindqvist, Christina, Anna Gudmundson, and Camilla Bardel. 2013. "A New Approach to Measuring Lexical Sophistication in L2 Oral Production." In *L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis*, edited by Camilla Bardel, Christina Lindqvist, and Batia Laufer, 109–26. European Second Language Association.
- Linnarud, Moira. 1986. *Lexis in Composition: A Performance Analysis of Swedish Learners' Written English*. Malmö: Liber Forlag Malmö.
- McCarthy, Michael, and Anne O'Keeffe. 2010. "What Are Corpora and How Have They Evolved?" In *The Routledge Handbook of Corpus Linguistics*, edited by Michael McCarthy and Anne O'Keeffe, 3–13. Abingdon, New York: Routledge.
- Milton, James. 2010. "The Development of Vocabulary Breadth across the CEFR Levels." In *Communicative Proficiency and Linguistic Development: Intersections Between SLA and Language Testing Research*, edited by Inge Bartning, Maisa Martin, and Ineke Vedder, 211–32. European Second Language Association.
- Morris, Lori, and Tom Cobb. 2004. "Vocabulary Profiles as Predictors of the Academic Performance of Teaching English as a Second Language Trainees." *System* 32 (1): 75–87. <https://doi.org/10.1016/j.system.2003.05.001>.
- Nation, Paul. 2006. "How Large a Vocabulary Is Needed for Reading and Listening?" *Canadian Modern Language Review* 63 (1): 59–82.
- . 2013. *Learning Vocabulary in Another Language*, 2nd ed. Cambridge: Cambridge University Press.
- . 2020. "The BNC/COCA Word Family Lists." Accessed September 1, 2019. https://www.wgtn.ac.nz/_data/assets/pdf_file/0005/1857641/about-bnc-coca-vocabulary-list.pdf.
- Nation, Paul, and Robert Waring. 1997. "Vocabulary Size, Text Coverage and Word Lists." In *Vocabulary: Description, Acquisition and Pedagogy*, edited by Norbert Schmitt and Michael McCarthy, 6–19. Cambridge: Cambridge University Press.
- Read, John, and Paul Nation. 2006. "An Investigation of the Lexical Dimension of the IELTS Speaking Test." *IELTS Research Reports* 6: 1–25.
- Scott, Mike. 2004. *Oxford WordSmith Tools* (version 4.0). Oxford: Oxford University Press.
- Sonomura, Marion Okawa. 1996. *Idiomacity in the Basic Writing of American English*. New York, NY: Peter Lang.
- Staples, Shelley, and Randi Reppen. 2016. "Understanding First-Year L2 Writing: A Lexico-Grammatical Analysis across L1s, Genres, and Language Ratings." *Journal of Second Language Writing* 32: 17–35. <https://doi.org/10.1016/j.jslw.2016.02.002>.
- Troia, Gary. 2007. "Research in Writing Instruction: What We Know and What We Need to Know." In *Shaping Literacy Achievement: Research We Have, Research We Need*, edited by Michael Pressley, Alison Billman, Kristen Perry, Kelly Reffitt, and Julia Reynolds, 129–56. New York: Guilford Press.
- Vuković Stamatović, Milica. 2019. "Vocabulary Complexity and Reading and Listening Comprehension of Various Physics Genres." *Corpus Linguistics and Linguistic Theory*. <https://doi.org/10.1515/cllt-2019-0022>.

- Webb, Stuart. 2008. "Receptive and Productive Vocabulary Sizes of L2 Learners." *Studies in Second Language Acquisition* 30 (1): 79–95. <https://doi.org/10.1017/S0272263108080042>.
- Wang, Jing, Shao-lan Liang, and Guang-chun Ge. 2008. "Establishment of a Medical Word List." *English for Specific Purposes* 27: 442–58.
- West, Michael. 1953. *A General Service List of English Words*. London: Longman.