Nordic Journal of Language Teaching and Learning

A Bigram-Based Exploration of Phraseological Development in Norwegian Secondary School Students' Writing in English L2

Kaja H. S. Ø. Evang

University of Oslo

kaja@evang.no

Abstract

This study investigates phraseological units produced by Norwegian secondary school students (aged 13 to 17) writing in English L2. The study employs association measure as a way of calculating the collocational strength and certainty of words in word pairs (bigrams) in the L2 learners' writing. The association measures MI and *t*-score have been shown to be reliable measures for telling learner language apart from native language. Durrant and Schmitt (2009) and Granger and Bestgen (2014) found a higher proportion of high MI-scoring bigrams to be a marker of more advanced language, native or nativelike, while a higher proportion of high *t*-scores was associated with less advanced language. The present study includes language from a lower proficiency level than has previously been investigated. The pattern of association measures found for intermediate Norwegian learners of English does not match the previous findings for advanced learners. Instead, an initial decline is uncovered, in that the students produce a higher proportion of bigrams with high MI scores and high *t*-scores in the first year than they do one and two years later. At higher levels of proficiency, the scores increase again, the pattern resembling previous findings. The article considers possible explanations and discusses applications for teaching.

Keywords

Young learner language, English L2, longitudinal, phraseology, collocation, association measure.

Nordic Journal of Language Teaching and Learning, Vol. 10, No. 2, 2022. ISSN: 2703-8629.
https://doi.org/10.46364/njltl.v10i2.1001

1. Introduction

This project is inspired by several years of teaching English in Norwegian lower secondary school (8th to 10th grade), where I regularly come across students who master the rules of spelling, sentence structure, and grammar – and who nevertheless produce texts that do not "sound English". What these students not yet master seems to be finding words that fit well together. To improve the phraseological competence of language learners, we need a way to describe their level of proficiency and identify realistic teaching goals. This study aims to explore, quantify, and measure the phraseological competence of Norwegian secondary school students' writing in English.

Words have friends. They prefer the company of some words to the company of others that appear to have equal characteristics. Knowing a word implies knowing its friends, or as J. R. Firth stated it: "You shall know a word by the company it keeps" (1957, p. 11). Already in 1983, Pawley and Syder observed how phraseology is an area where learner language differs markedly from native language. Yet, prefabricated stretches of language are central to language learning. Ortega (2013, p. 114) describes the process by which formulaic language may be learnt wholesale early on, and only later undergo analysis of internal structure. Several studies (Ellis, 2002, 2012; Jiang & Nekrasova, 2007; Conklin & Schmitt, 2008; Ellis et al., 2008; Ellis & Simpson-Vlach, 2009) show that both second language learners and native speakers of a language process formulaic language quicker than non-formulaic language, supporting the recommendation that learners be taught phraseological expressions.

To gain insight into the phraseological proficiency of Norwegian intermediate learners of English (students in 8th to 12th grade, from 13 to 17 years old), this study takes a statistical approach to the identification of phraseological units in texts written by students in secondary school. The study follows previous studies in operationalising association measures (*t*-score and MI score) of bigrams (two-word units) as a measure of phraseological performance, thereby providing a description of the proficiency and development of phraseology in the learners' writing. The methods of investigation are chosen to enable comparison with previous studies exploring the phraseological competence of learners of English at university level (Durrant & Schmitt, 2009; Granger & Bestgen, 2014; Bestgen & Granger, 2018).

The research questions are:

RQ1: Does the same pattern that has been shown for advanced learners, i.e. an increase in high MI scores and a decrease in high *t*-scores, hold for Norwegian intermediate learners of English as their level of proficiency increases?

RQ2: What additional insights can be gained into the phraseological proficiency of Norwegian intermediate learners of English by measuring the MI and *t*-score of bigrams in written texts?

The study is corpus driven, in the sense that it takes a bottom-up approach, and the main objective of RQ2 is to explore the data and see what patterns can be revealed. RQ1 aims to fill a knowledge gap by tracing the phraseological development of intermediate learners along the lines of investigations that have previously provided insights about advanced learners of English. Based on the answers to these questions, I hope to show that the results and knowledge gained in this study can be applied to teaching.

In Section 2, statistical phraseology and some important concepts are defined. Section 3 presents previous research, and material and method are accounted for in Section 4. The results are presented in Section 5. Section 6 revisits the research questions, summarises findings, and discusses some aspects applicable to teaching, before Section 7 rounds off with suggestions for further research.

2. A Statistical Approach to Phraseology

There are arguably two main traditions in phraseology. In the traditional, 'phraseological' approach, items of interest for phraseological research are defined by degree of idiomaticity. These degrees are described by Howarth (1998) as a continuum of relations between words, ranging from free combinations (e.g. *red shirt*), via restricted collocations (e.g. *red carpet*) and figurative idioms (e.g. *caught red-handed)*, to pure idioms. (e.g. *red herring)*. In this approach, native speaker intuition is traditionally called upon for determining if items count as phraseological units. More recently, an alternative approach to phraseology has developed, alternately referred to as the 'statistical', 'distributional', 'probabilistic', or 'frequency-based' approach. In this tradition, phraseological units are determined by statistical measures. In a collection of texts such as a corpus, computer software can count all the words and determine how likely they are to appear in the vicinity of each other. If they do so more often than chance would predict, the words *collocate*. Several types of phraseological units can be identified, depending on the research interest and the method of calculation. The words can be consecutive or with interrupting words, and in a certain order. Following the method of previous studies (Durrant & Schmitt, 2009; Granger & Bestgen, 2014; Bestgen & Granger, 2018), the items considered in the present study are *bigrams*, sets of two consecutive words.

When calculating the probability of words occurring together, the resulting number is referred to as an 'association measure' or an 'association score' (Gries, 2015). Words with a high probability of occurring together will have a high score, and hence constitute items of interest from a phraseological point of view. There are different ways of calculating such scores. The two measures used in Durrant and Schmitt (2009), and subsequently in Granger and Bestgen (2014), and Bestgen and Granger (2018), are *t*-score and MI score (Mutual Information). Others could have been considered, such as *z*-score, log-likelihood, and 'cubed' MI score (MI3). According to McEnery et al. (2006, p. 217), MI3 is particularly well suited for locating collocations for pedagogical purposes. For comparability with previous studies, the present study focusses on MI and *t*-score.

When calculating the MI score, the frequency of words appearing together is compared to the expected frequency of them occurring together by chance. If the score is close to 0, the two words have a low *collocational score*, and the co-occurrence may be random. According to Hunston (2002), MI score measures the amount of non-randomness present for co-occurring words, and this measure is considered significant when it exceeds a score of 3. MI score has been shown to identify collocations that native intuition recognises as idiomatic expressions, making it particularly useful for the study of phraseology (Ebeling & Hasselgård, 2015). The more "fixed" a phrase is, the higher the MI score, which is why MI score is said to measure strength of collocation (Hunston, 2002, p. 71).

While MI score measures strength of collocation, *t*-score measures certainty of collocation. By including the standard deviation in the calculation, the sample size is taken into account, and the more frequent a combination is, the higher *t*-score it receives. McEnery et al. explain that "[t]he score can be computed by subtracting the expected frequency from the observed frequency and then dividing the result by the standard deviation" (2006, p. 56). A *t*-score of minimum 2 is required for two words to be considered collocations. High *t*-score identifies collocations with a high frequency, which are often grammatical words, such as *of the* and *it was*. It can also identify more traditional idiomatic expressions, provided they have a high frequency in the material, such as *World War*.

The statistical approach to phraseology is useful for computational linguistics, such as natural language processing, and the development of reliable measures for automated proficiency scoring, the aim of Bestgen and Granger (2018). Another advantage of the statistical approach is that it is "objective" in the sense that any researcher employing the same method on the same material, will arrive at the same list of items with the same statistical scores and probabilities, making researching phraseology possible even without access to a native speaker's intuition.

This is the main motivation for choosing the statistical approach in the present study, along with the possibility of comparisons with previous studies.

One disadvantage of the statistical approach is that a "blind" measure of frequency will include sequences such as *and of the*, which intuition dismisses as a possible phraseological unit. Statistical measures are not enough to identify relevant phraseological units, as observed by Ellis et al. (2015). A possible response is a requirement of semantic unity, which the present study seeks by relying on word class combinations (i.e. noun + noun, adjective + noun, and adverb + adjective).

3. Previous Research

According to Sinclair (1991), much of language production is based on 'the idiom principle', where whole stretches of words are chosen at once, allowing for speed and fluency in language output because fewer decisions need to be made, resulting in language production where many words form part of larger, formulaic units. The idiom principle works alongside 'the open-choice principle', by which the language user combines words freely into grammatical utterances. Kjellmer (1991) explores how these two principles can explain some of the differences between learner and native language. He describes how learner language is marked by more pauses and suggests that this is due to learners having access to fewer automated expressions, relying more on the open-choice principle in language production, thus having to make decisions more often, as the learner's "building material is individual bricks rather than prefabricated sections". This also leads to language that "seem[s] contrived or downright unacceptable to native ears" (Kjellmer, 1991, p.p. 124-125).

Durrant and Schmitt (2009) react to this claim by Kjellmer (1991) that learners rely on singular words when acquiring a second language, and set out to investigate the presence, or absence, of formulaic language in advanced L2 learner output. They compare association scores of bigrams in English texts written by non-native university students to those in texts written by native English-speaking university students. The calculation of the association scores is based on the words' presence in the British National Corpus (BNC1994, original 1994 version)[1]. They find a higher proportion of rare combinations (below threshold, BT, i.e. less than 5 tokens in the BNC1994) in the native texts than in the learner texts. With regard to high *t*-scoring bigrams, they find that the learners produce about as many as the native speakers, but with a more limited repertoire, and a higher degree of repetition of a few trusted items (c.f. lexical and phraseological

[1] http://www.natcorp.ox.ac.uk/

'teddy bears', Hasselgren, 1994; Hasselgård, 2019). When comparing MI scores, Durrant and Schmitt (2009) find consistent underuse of bigrams with an MI score ≥7 by the L2 learners, compared to native writers. They conclude that while Kjellmer's observation that "there is something missing" (Durrant & Schmitt, 2009, p. 174) from learner writing is not wrong, this does not necessarily support the idea that adult L2 learners construct their writing from singular words alone. What is missing are not multi-word units altogether, but rather low-frequency collocations, characterised by a high MI score, while high-frequency collocations, identified by a high *t*-score, are used by learners just as much as by natives.

Granger and Bestgen (2014) build on Durrant and Schmitt's (2009) methodology and compare the writing of university students at different proficiency levels of English L2 writing, with a view to developing a reliable description of the characteristics of learner language levels for use in automated scoring. They had texts from the International Corpus for Learner English (ICLE) assessed and marked for proficiency level with reference to The Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001). Comparing the results from their two sub-corpora, one 'intermediate' (B) and one 'advanced' (C), Granger and Bestgen (2014) report "a smaller proportion of lower-frequency, but strongly associated, collocations [attested by MI] and a larger proportion of high-frequency collocations [attested by *t*-score] in the intermediate learner texts than in the advanced learner texts" (2014, p. 238). They find these results to hold true for both types and tokens of bigrams in most of the examined categories of bigrams. They argue that both the high-frequency collocations identified by high *t*-scores and the low-frequency, strongly associated collocations identified by high MI scores are important in learner writing, but that the balance shifts from a reliance on the former towards an increased reliance on the latter as the learners' proficiency increases. Just as Durrant and Schmitt (2009), Granger and Bestgen (2014) find more BT bigrams in the higher proficient group. They conclude that the same pattern is shown to hold between intermediate and advanced learners, as Durrant and Schmitt find between advanced learners and natives. They propose that the category "All", which contains bigrams of all the words the learners produce regardless of word class, is supremely suited for investigation from the point of view of automated scoring, since it shows the difference between intermediate and advanced learners most clearly and is more efficiently automatically retrievable. They round off by recommending that similar studies be carried out for learners of lower proficiency than CEFR B (i.e. CEFR A), and for learners in an immersion situation, in contrast to the relatively input-poor, foreign language environment of the ICLE learners.

In a 2018 study, Bestgen and Granger test their findings on a longitudinal corpus. They introduce the term "collgram" for *n*-grams with an assigned association score. This time they use texts from the LONGDALE (Longitudinal Database of Learner English) corpus, written by French-speaking undergraduate university students. Each student wrote two essays, one in their first year of university, and one in their third year. They were required to write their essays on the same topic both times, to enhance comparability between the two texts. Bestgen and Granger (2018) find that learners produce more non-collocational (both MI and *t*-score) collgrams in Year 1 than in Year 3, and that they produce more tokens with high MI scores in Year 3. They find the same tendency as in their study from 2014, and as Durrant and Schmitt (2009), that the proportion of high-scoring *t*-score collgrams is higher at the lower proficiency level (Year 1) than in Year 3. Again, they find a higher percentage of BT bigrams in Year 3, and again, although they do not investigate this further, they mention that this "clearly deserves careful investigation," (2018, p. 288). Bestgen and Granger (2018) compare their 2014 and 2018 studies to see whether anything is gained from using a longitudinal corpus rather than a pseudo-longitudinal one, which is otherwise more easily accessible. They conclude that the two studies give very similar results, without thereby granting that collecting longitudinal data is not worthwhile.

Most studies of phraseology in written learner language have explored the language of university students. Durrant and Schmitt (2009) compare L2 university students to natives, Granger and Bestgen (2014) compare different proficiency levels of university L2 students to each other, calling these levels "intermediate" and "advanced", and Bestgen and Granger (2018) compare the writing of university L2 students at different stages of development. Studies of learners with a lower proficiency in English have not been as abundant, and longitudinal studies are scarce. This leaves room for the present study, focussing on the language of secondary school students collected on separate occasions throughout their education. The study aims to gain knowledge about the phraseological proficiency and development of these L2 learners, knowledge that will be valuable for the identification of realistic teaching goals, assessment practices, and the development of teaching resources. Although Granger and Bestgen (2014) refer to their two groups of university students as "intermediate" and "advanced", these are all still more advanced than learners in secondary school. In the following, I will use "intermediate" to refer to the proficiency level of students in secondary school (8^{th}-12^{th} grade), and "advanced" to refer to the level of university students.

4. Material and Method

Texts from the TRAWL (Tracking Written Learner Language) corpus (Dirdal et al., 2022), a corpus of written learner texts currently under compilation in Norway, make up the main part of the material for this study. There are 890 texts from Norwegian secondary school, 734 of which are from a lower secondary school (8th to 10th grade) and 156 from an upper secondary (11th-12th grade), both schools in Eastern Norway. Included are also 316 texts from the Norwegian part of the ICLE corpus[2] (ICLE-NO) written by Norwegian undergraduate university students during the late 1990s and early 2000s, as well as 413 texts from the LOCNESS corpus[3], written by American students and British native speakers, 2/3 of whom were university students, and 1/3 A levels students. The LOCNESS texts were written in 1991 and 1995. The assembled material is assumed to constitute rising levels of proficiency, from the lowest in 8th grade, up through to the university students in ICLE-NO. The LOCNESS material, written by native English speakers, is taken to represent the highest level of proficiency.

The material from lower secondary school contains true longitudinal data. It was collected from two cohorts: 35 students born in 2001, and 47 students born in 2002. Where available, material from all three years of lower secondary school (8th, 9th, and 10th grade) was collected for each of these 82 students. In addition, 46 students from upper secondary school, born in 1998, took part in the study. There are texts from 46 students in 11th grade, as well as from 12th grade for seven of these. For a somewhat better balance in size between the sub-corpora of the material, and because the material from 12th grade is too scarce to be useful on its own, the texts from 11th and 12th grade are merged under the label "upper secondary school". The number of texts from each level is displayed in Table 1.

Table 1: Sub-corpora in this study

	8th grade	9th grade	10th grade	11th grade	12th grade	ICLE-NO	LOCNESS
Number of texts	240	266	228	131	25	316	413
Total number of words	158,133	191,782	163,365	98,026	19,154	211,418	325,583
Average number of words per text	659	721	717	748	766	669	788
Standard deviation	346	341	399	274	254	193	564
Trimmed mean	625	717	713	746	765	665	778

[2] https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html
[3] https://uclouvain.be/en/research-institutes/ilc/cecl/locness.html

From upper secondary school, ICLE, and LOCNESS the material is pseudo-longitudinal, as it does not contain material written by the same individuals. The texts in ICLE and LOCNESS are also considerably older, written 15-20 years before the material from secondary school. Coinciding with the great shift towards information technology and social media, this may have an impact on the comparability of vocabulary in the texts (see also discussion of suitability of BNC1994 as reference corpus in Section 6).

In the following, where not otherwise stated, the account of the texts describes the part of the material that comes from the TRAWL corpus (lower and upper secondary school). The label "text" is somewhat inaccurate. What is counted here as one text is in fact one document handed in by a student as an answer to an assignment. Quite often these assignments contain several tasks to be answered, so one document ("text") might be a collection of several, shorter texts. Since the genre of the texts is not controlled for, splitting the combined answers has not been considered worthwhile. Much larger sample sizes would have been required in order to control for genre and topic, since some of the word combinations are quite scarce in the learners' writing. In addition, the genres vary considerably across age levels, as the level of maturity and development of the students impose restraints on the types of tasks it is possible to assign to them. From 8th and 9th grade, the material contains a mix of genres, including informational texts, argumentative texts, reflective texts, and creative genres such as stories and diaries. From 10th grade and onwards, the material consists almost exclusively of argumentative and reflective writing. Task topics vary equally, a fact hard to avoid in a longitudinal corpus, short of asking students to answer the same assignment twice, as was done by the collectors of the LONGDALE corpus used by Bestgen and Granger (2018). Topics covered include the students' interests and activities, school-related topics such as bullying, school uniforms, and reading books, more general questions in society, such as social media, mobile phones, or immigration and integration, and topics specifically connected to the subject of English, such as the British Empire, English as a Global Language, and different English-speaking countries. For a more comprehensive account of the genres in the TRAWL material, see Hasund (2022). The genre and topic of the writing assignments can have a great impact on the vocabulary in the texts (Biber 2012), and as they are not rigidly controlled for, they can act as confounding variables in the interpretation of the results in the study. Studies of learner writing controlling for genre would be very welcome in the future.

The collected texts span 12 writing prompts in 8th grade, 13 in 9th grade, 10 in 10th grade, 10 in 11th grade, and 5 in 12th grade. In lower secondary school, each student typically answered 4-5 of the writing prompts each school year. In the TRAWL corpus (Dirdal et al., 2022), the

assignments are assigned a 4-letter code. The assignment codes and the number of answers are displayed in Table 2. The highest number of students who answered the same assignment is 60, the lowest is one, while on average, there are 21 different answers for each prompt in lower secondary school. In upper secondary school, an average of 13 students answered the assignments in 11th grade, and 5 students on average in 12th grade.

Table 2: List of assignment codes with number of answers

8th grade		9th grade		10th grade		11th grade		12th grade	
Assign. code	# of answers	Assign. code	# of answers	Assign. code	# of answers	Assign. code	# of answers	Assign. code	# of answers
BAMA	22	FUFR	35	AMPR	4	AMER	24	CRAS	4
CCDD	2	FUFT	2	ARWO	60	ARWO	23	GLCH	6
CCSH	32	JIPS	35	CSCC	23	BRLA	1	LEED	2
FAPI	5	JIPT	2	ENGL	23	CHAR	12	MUSO	7
INDG	9	JITC	28	GUMP	24	ESWO	4	REGL	6
LOND	6	PLGE	12	LETO	1	HALV	11		
MPFF	34	SIXI	36	LETR	4	KURS	1		
MPFT	2	SIXT	1	ONOF	36	MALA	5		
RASS	37	TALC	6	WAPE	21	POWE	45		
TAUS	28	TBSM	23	WARP	32	WRIT	4		
TEEN	36	TLEE	31						
TTNY	27	WIWE	20						
		WIWN	35						

The students in the study all have a Norwegian language background, from Eastern Norway. Several studies (e.g. Altenberg & Granger, 2001; Paquot 2013, 2014) show L1 background to have an impact on the vocabulary the learners produce. Approximately 10% of the originally available material is excluded from the study, having been written by students whose first language is not Norwegian.

Following Durrant and Schmitt (2009), Granger and Bestgen (2014), and Bestgen and Granger (2018), the 1994 version of the British National Corpus (BNC1994) serves the purpose of a target language norm, in so far as the association scores (*t*- and MI score) are calculated for bigrams (word pairs) based on their presence in the BNC1994. The adequacy of the BNC1994 as a reference corpus may be questioned (see discussion in Section 6). It is chosen partly because it matches the methodology of previous studies, and partly because the web interface at http://bncweb.lancs.ac.uk/ offers a calculator where association scores can easily be retrieved. The 2014 version of the BNC had not yet been released at the time of study. The anonymised texts were POS-tagged (part-of-speech tagged) with the automatic tagger CLAWS7 (Constituent

Likelihood Automatic Word-tagging System)[4], to enable semi-automatic retrieval of bigrams based on word class. The corpus analysis software AntConc 3.5.8 (Anthony, 2019) was used to extract the bigrams with the Clusters/N-grams tool. Search Term was set to 'Words', and Cluster Size to 'Min 2/Max 2'. Results were sorted by range. For analysis of association measures, a range threshold of five learner texts was introduced, so that only bigram types found in at least 5 different texts were assigned MI and *t*-scores from the BNC1994. This was done to avoid a single student skewing the results, to keep the number of bigrams manageable, and to follow the methodology introduced by Durrant and Schmitt (2009).

The search strings used to extract the bigrams were *_JJ *_NN for adjective+noun, and *_NN *_NN for noun+noun. The CLAWS7 tagger differentiates degree adverbs and general adverbs, and tags *very* as the former and *really* as the latter. As these two words are frequently used by learners, the search strings *_RG *_JJ for degree adverb+adjective, and *_RR *_JJ for general adverb+adjective were both included. The adverb *really* could possibly better be tagged a degree adverb, as it is used emphatically in most cases by the learners, e.g. *really great*. The category "All", included in the studies by Granger and Bestgen (2014) and Bestgen and Granger (2018), was not possible to include with the semi-manual extraction method used in this study. As the "All" category includes many grammatical words and combinations that do not constitute semantic units, they are deemed less interesting for the present study, where the aim is to inform teaching, rather than automated processing.

The number of bigrams, their frequencies, range, and collocational scores were registered in Microsoft Excel, which was also used to conduct calculations. The sophisticated statistical tools employed by Bestgen and Granger (2018) were not available in this study, which relies on simple statistics. Relative frequencies and type-token ratios have been calculated for each learner level as a whole, without regard for individual learner trajectories. For this reason, the study is not immune from the possibility of 'Simpson's paradox' – a statistical phenomenon that makes a comparison reversible when several observations are grouped (Moore et al., 2021). This must be kept in mind when interpreting the results of this study, and further studies using more sophisticated statistical methods will be very welcome.

Following Granger and Bestgen (2014), the collocational scores of the bigrams are grouped into the four categories "Non collocational" (NC), "Low collocational" (L), "Medium collocational" (M), and "High collocational" (H), both for MI scores and for *t*-scores, and a Below Threshold (BT) category, which is bigrams with fewer than 5 tokens in the BNC1994,

[4] I am grateful to Signe Oksefjell Ebeling at the University of Oslo for helping me with the tagging of the texts.

deemed too few to assign a reliable association measure[5]. The BT bigrams are included in the statistics for comparison of proportions across learner levels and discussed further in Section 6. The MI and *t*-score intervals for each category are listed in Table 3.

Table 3: Intervals for categories of MI and t-*score (Granger and Bestgen 2014)*

	MI score	*t*-score
Non collocational (NC)	< 3.00	< 2.00
Low collocational (L)	≥ 3.00 and < 5	≥ 2.00 and < 6.00
Medium collocational (M)	≥ 5.00 and < 7	≥ 6.00 and < 10.00
High collocational (H)	≥ 7.00	≥ 10.00

The major parts of the analysis focus on bigram types, ignoring tokens, and compare the percentages of the different categories of collocational scores as produced by the different learner levels. Bigram types are seen as more relevant than tokens, as the aim of this study is to inform teaching, and the repertoire of bigram types accessible to learners at different developmental stages is arguably more interesting than the number of times they use each one. While this method does not replicate the level of consideration for individual variability which the studies of Granger and Bestgen (2014), and Bestgen and Granger (2018) commendably do, concentrating on what is typical and common for several learners is deemed most urgent in the present study. Ebeling and Hasselgård (2015, pp. 211-212) warn that frequency alone cannot tell whether a learner has acquired nativelike command of a lexical word or phrase, as appropriate use also includes applying it in the correct context. They recommend inspecting the search results by concordance lines, KWIC (Keyword in Context). To this end, checks of concordance lines have been conducted on a random sample of the bigrams, as has checks of concordance plots illustrating dispersion across texts, to substantiate that the distribution of bigrams is balanced, and to find examples.

The main part of this study is done on aggregated data. However, as a spot check to assess the results of the main study, the output from one randomly chosen student in lower secondary school has been measured in detail across three years. The main takeaway from that undertaking is that there is a considerably higher proportion of BT bigrams in that material. The remaining above threshold bigrams do not contradict the general trend seen in the aggregate data, however,

[5] Note that this is a different threshold of 5 than the one used to determine which bigrams to include in the investigation. The first threshold was that of five student texts, the minimum distribution required for bigrams to be included in the study.

conclusions cannot be drawn on this limited evidence. A more detailed analysis of the material from the single student can be found in Evang (2019).

5. Results

5.1 Overall Results

Overall, 3.9% of the bigram types in the relevant word classes have been included for study, the remaining 96.1% falling below the 5-range threshold. However, these 3.9% represent 26.7% of the tokens, which means that just over a quarter of the bigram tokens produced by these students are examined in this study. Table 4 displays a general word count and type-token measures. The numbers do not differ greatly between the proficiency levels, with two exceptions. The smaller sample size is probably the cause of the higher type-token count for upper secondary school (6,403), as the sample size is known to impact type-token measures (Cobb & Horst, 2015, p. 192). LOCNESS, on the other hand, has the largest sample size, yet a fairly high type-token measure. This may be indicative of the native students possessing a wider vocabulary than the Norwegian learners.

Table 4: Word types, tokens and types per 100,000 words

Sub-corpora	Number of words (tokens)	Word types	Types per 100,000 words
8th grade	158,133	7,295	4,613
9th grade	191,782	8,388	4,374
10th grade	163,365	6,844	4,189
Upper secondary	117,180	7,503	6,403
ICLE-NO	211,418	10,262	4,854
LOCNESS	325,583	17,000	5,221

The total number of bigram types from each level is displayed in Table 5, along with the number of bigrams that are above a minimum range threshold of 5 different texts. The bigrams are sorted according to which word class (bigram category) they belong to. To enable comparability between learner levels, Table 5 also shows the frequencies normalised per 100,000 words.

Table 5 shows that the bigram type adjective+noun (JJ NN) is by far the most numerous, followed by the noun+noun type (NN NN). The degree adverb+adjective group (RG JJ) follows, while the general adverb+adjective group (RR JJ) counts the fewest bigrams. There is a considerable difference in how many of these bigrams the different levels produced. The number of JJ NN bigrams increases sharply, from just over 1,500 per 100,000 words in lower secondary school, to around 2,500 in upper secondary school, ICLE-NO, and LOCNESS. The number of

NN NN bigrams fluctuates more, with the lowest number in 10th grade at 657 per 100,000 words, to the highest in LOCNESS at 1,109. The number of RR JJ bigrams remains generally low across all learner groups, though there is an increase from 164 per 100,000 words in 8th grade to 217 in upper secondary school and 252 in LOCNESS. In the RG JJ group, the total number of bigrams fluctuates, with the lowest number in LOCNESS at 124 bigrams per 100,000 words, and the highest in upper secondary school at 209.

However, the number of RG JJ bigrams with a minimum range of 5 texts decreases steadily, from 25 per 100,000 words in 8th grade to only 6 in ICLE-NO, and 7 and LOCNESS. A possible explanation for this decrease could be the types of expressions these bigrams represent. The adverb part is in most cases *very* or *so*, e.g. *so happy*, *very bad*, *so angry*, *very beautiful*. Quite a few of them are mainly associated with spoken language, such as *so cool*, *very cool*, and *so excited*, which are exclusively found in 8th and 9th grade. In the BNC1994, *so excited* has a frequency of 2.02 per million words in spoken language, and only 1.29 in written language. Examples of the few found in LOCNESS are *very significant*, *very likely*, *as important*. As students reach higher levels of proficiency, they learn to avoid expressions typically associated with speech in their writing. Some of the writing prompts offered in 8th and 9th grade, such as diaries and personal letters, could also contribute to this difference.

Table 5. Number of bigram types ranging ≥5 texts and in total, absolute and normalised per 100,000 words

		JJ NN		NN NN		RG JJ		RR JJ	
Sub-corpora		≥5	Total	≥5	Total	≥5	Total	≥5	Total
8th grade	Absolute	64	2,505	21	1,402	39	323	6	260
	Normalised	40	1,584	13	887	25	204	4	164
9th grade	Absolute	109	3,038	14	1,467	33	363	11	352
	Normalised	57	1,584	7	765	17	189	6	184
10th grade	Absolute	138	2,525	40	1,073	19	322	7	307
	Normalised	84	1,546	24	657	12	197	4	188
Upper secondary	Absolute	86	2,812	12	1,059	11	245	3	254
	Normalised	73	2,400	10	904	9	209	3	217
ICLE-NO	Absolute	127	5,385	23	1,447	13	412	7	455
	Normalised	60	2,547	11	684	6	195	3	215
LOCNESS	Absolute	143	9,209	34	3,611	22	405	4	822
	Normalised	44	2,828	10	1,109	7	124	1	252

In the following sections, only bigram types ranging from at least five texts are considered. For brevity, only the word 'bigram(s)' is used. However, it should be kept in mind that this refers to bigram *types* (not tokens). Note that this is a different threshold than the 5-token limit for assigning a collocational score. Bigrams with less than 5 tokens in the BNC1994 are included in the study and counted in the following as the category 'below threshold' (BT).

5.2 Adjective + Noun (JJ NN)

The students in 8th grade produced 64 different bigrams of this category, the 9th graders 109, the 10th graders 138, the upper secondary school students 86, the students in ICLE-NO 127, and in LOCNESS 143. The total number of tokens can be seen in Table 5.

Figure 1 displays the percentage of the JJ NN bigrams that fall into the different categories of collocational strength (Non collocational - NC, Low - L, Medium - M, and High - H), comparing each learner level side-by-side in what is assumed to be an increasing order of proficiency. The below threshold (BT) percentage is of course the same for MI and *t*-score. The proportion of M and H MI scores falls from 8th to 9th grade, before it rises again. It is highest in LOCNESS. Equally, the proportion of NC and L MI scores rises from 8th to 9th grade, where it peaks, before it falls gradually, reaching a low point in LOCNESS. Figure 2 displays the categories of *t*-scores in the same way. The pattern is similar, with the low point now in 10th grade, but the proportion of H *t*-scores is slightly lower in LOCNESS (native students) than in ICLE-NO. This pattern is to be expected. As Durrant and Schmitt (2009) found, learners tend to produce a higher

Figure 1. JJ NN bigrams, MI scores.

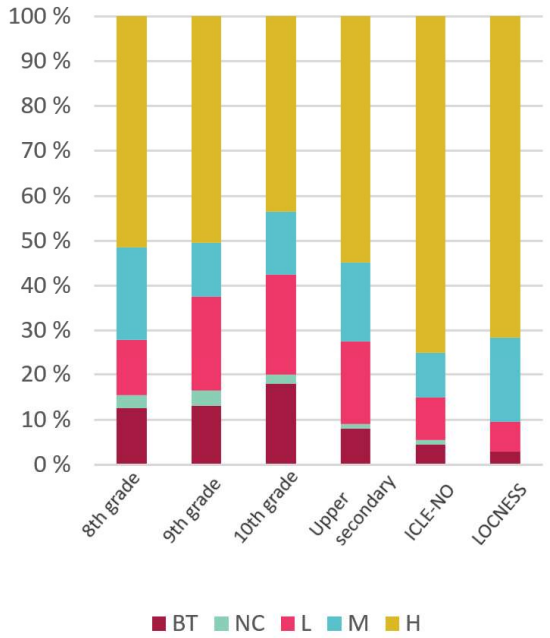

Figure 2. JJ NN bigrams, *t*-scores.

proportion of bigrams with a high *t*-score, while natives produce a higher proportion of bigrams with a high MI score.

Figures 3-8 show the intersection between the association scores. Observable here is the paucity in lower secondary school, particularly in 9th and 10th grade, of bigrams with both a high MI score and a high *t*-score. These bigrams, which make up 26% of the bigrams in the adjective+noun category in LOCNESS, only add up to 6% and 9% in 9th and 10th grade. Examples of bigrams with both a high MI score and a high *t*-score in LOCNESS are *great deal*, *short term*, *long term*, *certain aspects*, *common sense*, *near future*, *medical profession*, *working class*, *human race*, *European unity*, *Catholic church*, *national lottery*. Examples of the few from 9th grade are *little bit*, *brown hair*, *human beings*, *pop music*. Apart from the drop from 8th to 9th grade, these observations support the view that association scores increase as the learners become more proficient.

MI scores	NC	L	M	H
H	0 %	0 %	3 %	11 %
M	0 %	5 %	6 %	28 %
L	0 %	5 %	6 %	13 %
NC	3 %	3 %	5 %	0 %

t-scores

Figure 3. Distribution of 8th grade JJ NN bigrams. Intersection of MI and t-*scores.*

MI scores	NC	L	M	H
H	0 %	4 %	1 %	6 %
M	0 %	4 %	4 %	19 %
L	1 %	7 %	4 %	21 %
NC	3 %	6 %	4 %	4 %

t-scores

Figure 4. Distribution of 9th grade JJ NN bigrams. Intersection of MI and t-*scores.*

MI scores	NC	L	M	H
H	0 %	2 %	1 %	9 %
M	0 %	7 %	6 %	20 %
L	0 %	8 %	3 %	14 %
NC	2 %	6 %	4 %	1 %

t-scores

Figure 5. Distribution of 10th grade JJ NN bigrams. Intersection of MI and t-*scores.*

MI scores	NC	L	M	H
H	0 %	2 %	0 %	16 %
M	0 %	5 %	8 %	20 %
L	0 %	5 %	7 %	17 %
NC	1 %	7 %	2 %	1 %

t-scores

Figure 6. Distribution of upper secondary JJ NN bigrams. Intersection of MI and t-*scores.*

MI scores	NC	L	M	H
H	0 %	1 %	3 %	16 %
M	0 %	5 %	1 %	43 %
L	0 %	1 %	4 %	15 %
NC	1 %	3 %	2 %	2 %

t-scores

Figure 7. Distribution of ICLE-NO JJ NN bigrams. Intersection of MI and t-*scores.*

MI scores	NC	L	M	H
H	0 %	1 %	4 %	26 %
M	0 %	3 %	7 %	35 %
L	0 %	1 %	6 %	11 %
NC	0 %	1 %	2 %	0 %

t-scores

Figure 8. Distribution of LOCNESS JJ NN bigrams. Intersection of MI and t-*scores.*

5.3 Noun + Noun (NN NN)

There are fewer bigrams in this category than in the JJ NN category. 8th grade produced 21 types of these noun+noun bigrams, 9th grade only 14, 10th grade 40, upper secondary school 12, the students in ICLE-NO 23, and in LOCNESS 34. The pattern of MI scores in Figure 9 resembles the one for JJ NN bigrams in Figure 1, in that the scores are lowest in 9th grade, but here they are highest in upper secondary school. The pattern of *t*-scores displayed in Figure 10 is similar to Figure 9.

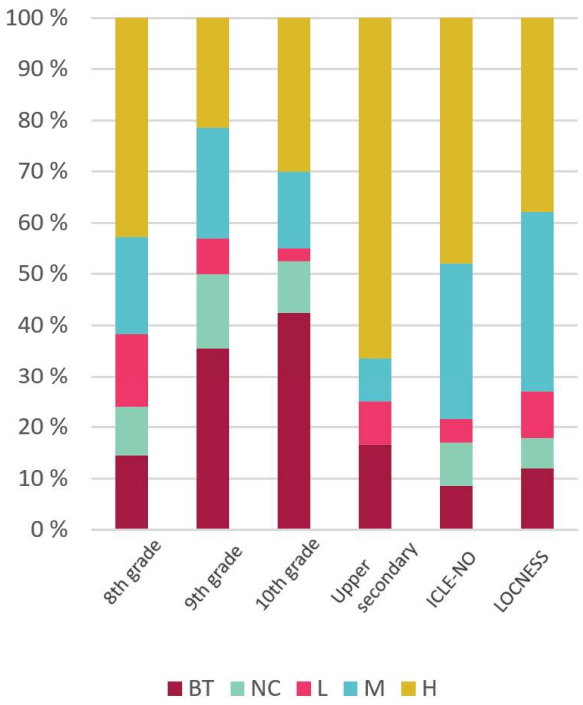

Figure 9. NN NN MI scores.

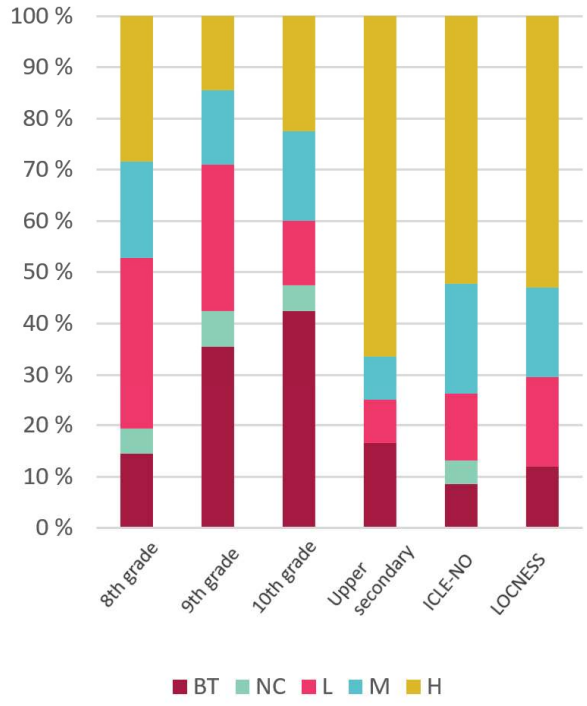

Figure 10. NN NN *t*-scores.

The intersection (Figures 11-16) shows that, while high MI scores and high *t*-scores are prevalent in the NN NN category, it is only in the higher levels of proficiency, i.e. upper secondary school, ICLE-NO, and LOCNESS (Figures 14-16), where most bigrams have both a high MI score and a high *t*-score. In lower secondary school (Figures 11-13), bigrams can have a high MI or a high *t*-score, but rarely both. Examples: *cartoon character* (8th grade, high MI, low *t*), *music industry*, *body language* (10th grade, medium MI, high *t*).

In the NN NN category, the scores fall from 8th to 9th grade, before rising again. The scores fall from upper secondary school to ICLE-NO and LOCNESS.

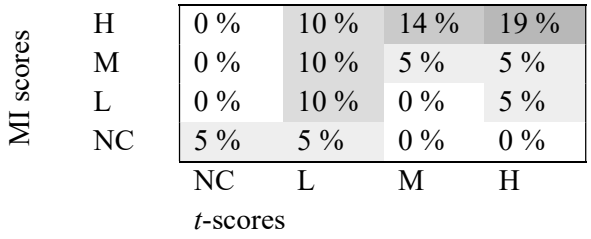

Figure 11. Distribution of 8th grade NN NN bigrams. Intersection of MI and t-*scores.*

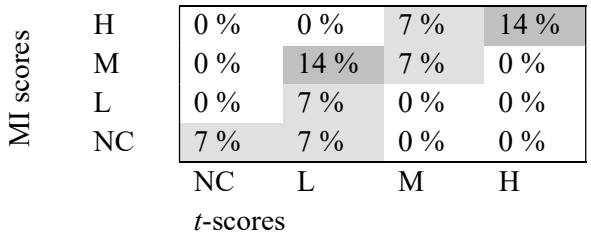

Figure 12. Distribution of 9th grade NN NN bigrams. Intersection of MI and t-*scores.*

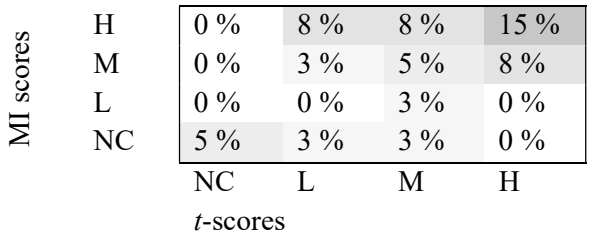

Figure 13. Distribution of 10th grade NN NN bigrams. Intersection of MI and t-*scores.*

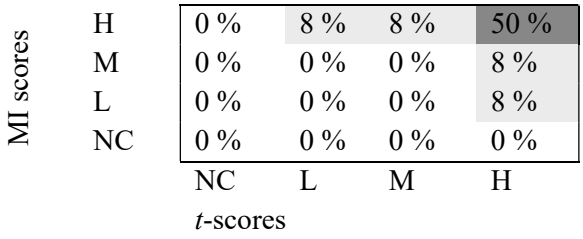

Figure 14. Distribution of upper secondary NN NN bigrams. Intersection of MI and t-*scores.*

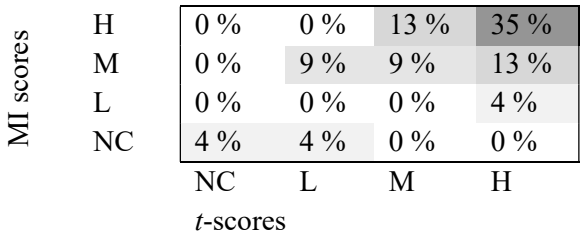

Figure 15. Distribution of ICLE-NO NN NN bigrams. Intersection of MI and t-*scores.*

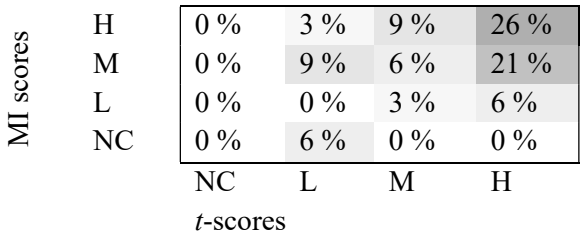

Figure 16. Distribution of LOCNESS NN NN bigrams. Intersection of MI and t-*scores.*

5.4 Degree Adverb + Adjective (RG JJ)

There is a steady decline in how many of these bigrams the informants use, which may be because many of these are expressions associated with spoken language that is avoided in the more proficient levels. None of these bigrams has a high MI score, probably because of the degree adverb typically used; a bigram involving such high frequency words as *as*, *so*, and *very* can never hope to be as exclusively associated as a high MI score requires. High t-scores, however, are prevailing, increasing, and, in LOCNESS, exclusive. The distribution is shown in Figures 17-18.

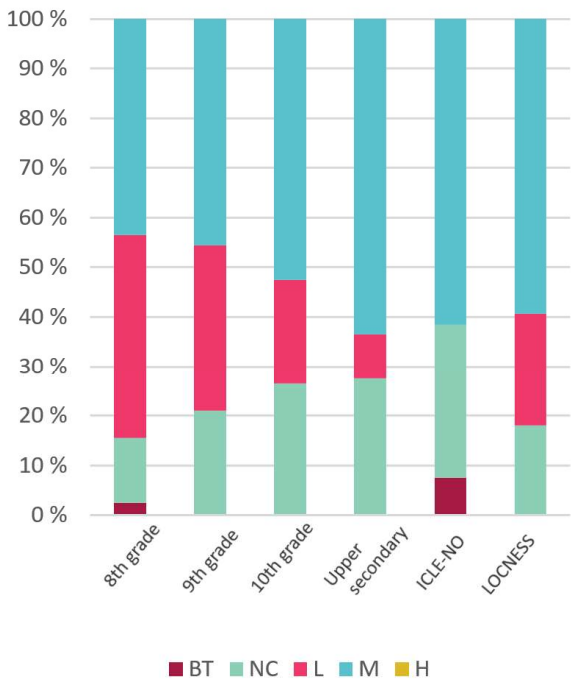

Figure 17. RG JJ bigrams, MI scores.

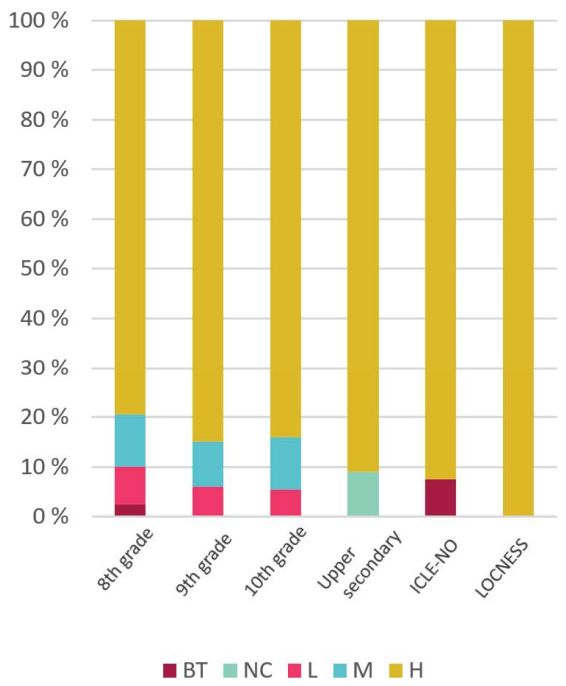

Figure 18. RG JJ bigrams, *t*-scores.

There are only two BT bigrams in this group: *so fun* in 8th grade and *too theoretical* in ICLE-NO. *So fun* appears to be an L1 transfer from the Norwegian "så gøy", while *too theoretical* appears to be triggered by the writing prompt (Example 1). Examples 2 and 3 show how the expression is used in two students' texts (my emphasis).

> **Example 1. Students were asked to discuss:**
> Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value. **(**ICLE writing prompt no. 3)**.**[6]
>
> **Example 2. Student answer.**
> The subjects in the teaching degree are relevant for the real world. The question is if they are **too theoretical**. Some of the subjects definitely are. (ICLE-NO-HO-0011.1)
>
> **Example 3**. **Student answer.**
> (…) the students having been bored to death by a **too theoretical** approach in earlier schooling, leading them to lose all interest in an area they have come to see as irrelevant (…). (ICLE-NO-HO-0008.1)

The expression *too theoretical* appears 13 times in ICLE-NO, and only 3 times in the BNC1994. The topic is possibly more closely associated with the context of writing in an institution of

[6] https://uclouvain.be/en/research-institutes/ilc/cecl/corpus-collection-guidelines.html

learning than in a general reference corpus such as the BNC1994. As an additional observation, the two examples illustrate language associated with less formal writing situations. These texts, along with others from the ICLE corpus, form the basis of Gilquin and Paquot's 2008 study on register in learner academic writing titled "Too chatty".

5.5 General Adverb + Adjective (RR JJ)

The distribution of scores in this category can be seen in Figures 19-20; however, since the number of items is considerably lower than in the other categories, this distribution must be treated with caution. In LOCNESS, the native speaker source in the material, there are four of these bigrams. They stand apart from the bigrams in the learner corpora, in that they have a more specific meaning and a higher MI score. The four bigrams in LOCNESS are *morally wrong*, *readily available,* and *sexually transmitted* (high MI score) and *ever increasing* (medium MI score).

In the learner corpora, including the university students in ICLE-NO, the RR JJ bigrams all involve emphatic use of the adverb: *also important*, *really good*, *completely different*, *really cool* (only 8th grade), and *extremely important* (upper secondary). There is little difference in how the learners use these RR JJ bigrams compared to the RG JJ bigrams with *very* and *so*. Adverbs like *really*, *also*, and *completely* are tagged by CLAWS7 as RR but should possibly rather be tagged as RG-adverbs when they are used as modifiers of adjectives.

As shown in Figures 19-20, high MI-scoring RR JJ bigrams show up first in 10th grade, and from then on steadily gain terrain – 14% in 10th grade, 33% in upper secondary school, 43% in ICLE-NO, and 75% in LOCNESS. BT and NC bigrams are only found in 9th and 10th grade, e.g., *really fun*, *relatively big*. 8th grade only has low and medium scoring bigrams of this category, all of them combinations with *really*, such as *really bad*, *really important*, *really nice*. Some of the bigrams do not constitute semantic units but are artefacts of the statistical approach: *long brown [hair]*, *both online [and offline]*, and *both good [and bad]* (the latter with a negative score, meaning that the words shun each other).

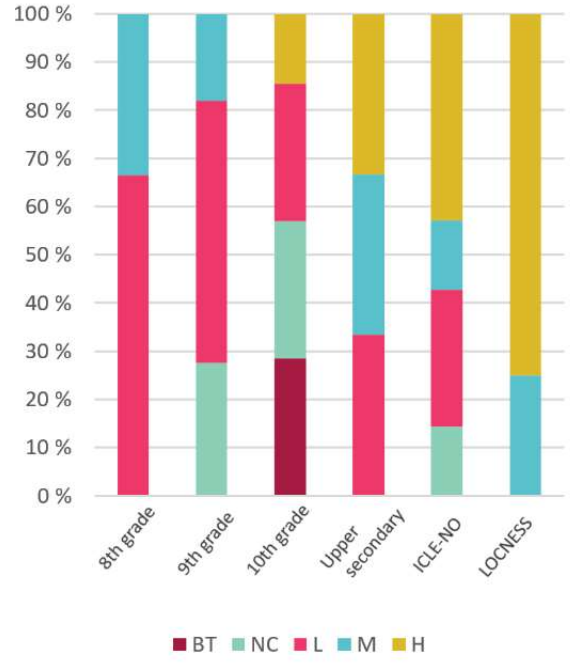

Figure 19. RR JJ bigrams, MI scores.

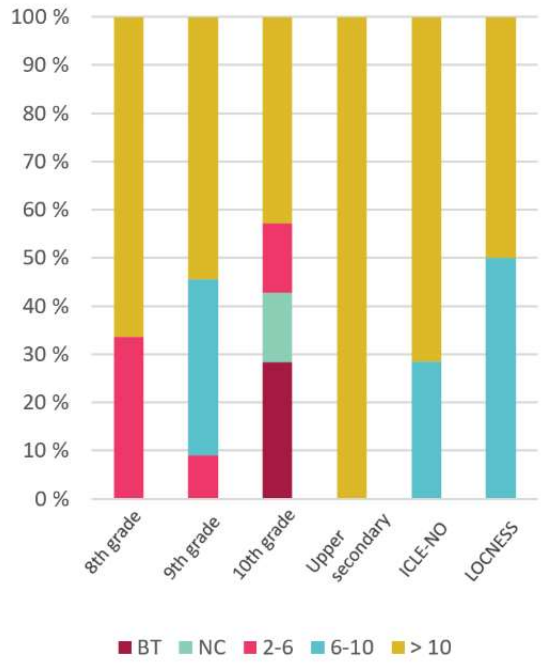

Figure 20. RR JJ bigrams, *t*-scores.

5.6 Impact of Writing Prompts

To assess the impact of the writing prompts, and to test a hypothesis that the increase of BT bigrams in 10th grade is related to the writing prompts, all ≥ 5-text bigrams from lower secondary school have been assessed for whether they are triggered by the wording in the writing prompt (WP) or not. In the JJ NN category, 27% of the bigrams in 8th grade are WP related, 30% in 9th grade, and 43% in 10th grade. In the NN NN category, 43% of the bigrams are WP related in 8th and 9th grade, and 60% in 10th grade. The intersect of the BT bigrams and the bigrams related to the writing prompts reveals that for the JJ NN category, 5 of 8 BT bigrams in 8th grade are WP related, 6 of 14 in 9th grade, and 20 of 25 in 10th grade. In the NN NN category, 2 of 3 BT bigrams in 8th grade are WP related, in 9th grade 3 of 5, and in 10th grade 13 of 17. This investigation shows that the increase in BT bigrams in 10th grade is to a large extent made up of bigrams triggered by the writing prompts. This is likely connected to the nature of these writing prompts, as in 10th grade, the students are increasingly asked to relate to and interact with texts provided in a preparatory phase or in the writing prompt itself. Hardly any of the bigrams in the RR JJ and RG JJ categories are BT nor related to the writing prompts. A similar inspection of the large amount of BT bigrams in the writing of the single student shows that these bigrams are not writing prompt related, with 0 instances in most grades/bigram categories, the largest number being 4 WP related NN NN bigrams in 10th grade, out of 27 BT bigrams in that category.

6. Discussion

The aim of this study is to gain knowledge about the phraseological development of Norwegian intermediate L2 learners of English, by comparing their development to what has previously been found for advanced L2 learners, and to explore parts of the newly collected TRAWL material with a view to identifying interesting aspects applicable to teaching of phraseology to young L2 learners. In this section the research questions are revisited, and some findings relevant for teaching are discussed.

6.1 Comparison to Previous Studies

> RQ1. Does the same pattern that has been shown for advanced learners, i.e. an increase in high MI scores and a decrease in high *t*-scores, hold for Norwegian intermediate learners of English as their level of proficiency increases?

Durrant and Schmitt (2009), Granger and Bestgen (2014), and Bestgen and Granger (2018) find that learners produce more bigrams with a high MI score and a smaller proportion of bigrams with a high *t*-score as they reach higher levels of proficiency. However, they all investigate the output from advanced learners of English. While the exact same methods as employed by these studies have not been available for this study, the results can be compared to some extent. The answer to research question 1 is 'no'. The same pattern has not been found for the intermediate learners in the TRAWL material, as has been found for the advanced learners. On the contrary, from 8th grade the pattern shows a decline in the proportion of high MI scores, as well as high *t*-scores. The scores reach a low point around 10th grade, from where they pick up again, to some extent MI scores, but especially *t*-scores increasing as the learners' level of proficiency advances. At this point, the learners rely on the common collocations rather than on the strongly associated ones. Only in the highest proficiency levels represented in this study do the *t*-scores dip down again, consistent with the findings in Durrant and Schmitt (2009), Granger and Bestgen (2014), and Bestgen and Granger (2018).

The decline in scores from 8th to 9th grade may be connected to the lower overall number of bigrams produced in 8th grade. It is also possible that the students in 8th grade rely more on the source material, and that 9th and 10th graders are more experimenting in their writing, akin to the way language learners may learn expressions wholesale early on and use them correctly before they begin deconstructing them (Cook 2008). With the method employed in the present study, we cannot dismiss that the pattern may be generated by Simpson's paradox (Moore et al. 2021)

and that a closer inspection of each individual learner's trajectory would reveal a different pattern.

6.2 Additional Insights and Ideas for Teaching

> RQ2. What additional insights can be gained into the phraseological proficiency of Norwegian intermediate learners of English by measuring the MI and *t*-score of bigrams in written texts?

Several characteristics of learner writing are demonstrated in the study. Particular attention will be paid here to the topic of general modifiers, where a teaching idea applying association scores is presented, and to the category of bigrams below threshold for assigning an association score (BT).

6.2.1 The Case of Very*: How to Avoid General Modifiers*

Learners have previously been shown to rely on a limited set of all-purpose modifiers (e.g. de Haan & van der Hagen, 2013), resulting in an overrepresentation of said modifiers in learner writing. This is also evident in the material in the present study, where learners are shown to rely heavily on adverbs such as *very*, *really*, and *so* and adjectives such as *good*, *big*, and *great* in adverb+adjective bigrams. This situation does not seem to improve considerably throughout the period of L2 development covered by this study, apart from the decrease in expressions associated with spoken language.

To remedy learners' limited repertoire of modifiers, association measures can be used first to identify expressions that include these general modifiers (e.g. bigrams with a low MI score and a high or medium *t*-score), and then to locate alternative expressions. The modifier *very* is frequently represented in learner texts, and the use of this modifier in 10th grade may serve as an example of how association measures can be applied in teaching, to help students vary their vocabulary. Of the 19 RG JJ bigrams in 10th grade, 12 begin with *very*. Two of these are taken directly from the assignment texts (*very soft* and *very simple*), and *very big* is discussed in the next paragraph. For the nine remaining generalized expressions with *very*, I suggest two approaches for finding more precise alternatives. One approach is to replace the entire expression with a more precise adjective, by locating synonyms or near-synonyms in a thesaurus. For this approach, I have used Thesaurus.com[7]. The other approach is to find a replacement for the

[7] http://thesaurus.com

modifying adverb (*very*) in the reference corpus by looking at association scores. My suggestions for alternative expressions resulting from these two approaches are presented in Table 6.

Of the 7 instances of *very big* in 10th grade, one modifies *house*, one modifies *city*, one modifies *country*, and 4 modify *differences*. The adjective *big* is in itself general and not very precise. Higher scoring alternatives involve replacing the adjective by looking at scores for JJ NN bigrams with the modified noun. As a replacement for *big differences* (MI score 3.44), higher scoring alternatives are *marked differences* (8.56), *significant differences* (8.54), and *striking differences* (7.22). When *big* modifies *house*, *city*, or *country*, however, no alternative suggests itself.

Bestgen and Granger (2018) state that a higher association score is equivalent to a better phraseological expression. They hold that the more closely associated (i.e. the higher the association score), the better a collgram is in terms of phraseological proficiency:

> Collgrams are particularly well suited for the analysis of L2 productions, as the different association scores enable a representation of phraseological proficiency as a continuum, from the best, most closely associated units to the downright incorrect ones, through a range of intermediary stages – good, weak and dubious.
> (Bestgen & Granger 2018, p. 280)

While this may generally be the case in most situations, the statement needs some modification. What constitutes a 'better' expression is highly dependent on context. Replacing these *very*-bigrams with one of the expressions suggested in Table 6 could easily lead to language that sounds stilted and pretentious. Register and language variety must be considered. *Jolly good* does not suit the same registers as *very good*. And while *damn good* might have a higher MI score than *very good*, it is hardly advisable as a replacement in (semi-)formal writing. Similarly, although *extremely common* has a higher MI score than *very common*, young students need probably not be encouraged to use expressions that are more hyperbolic and polarised than they already do. For this reason, teaching more moderate replacements should be considered, such as the expression *slightly different*, which I have listed in parenthesis as a suggested replacement for *very different*. A final caution: while teaching learners to vary their vocabulary is advisable, variation is not a goal in itself, but rather a means to greater precision or aesthetic value. Register awareness will not always result in variation in expression, as in some genres, such as scientific writing, the requirement of precision leads to lower variability in expression (Alley 2018).

Table 6: Suggested replacements for bigrams with very*, 10th grade RG JJ*

Original bigram	Adjective replacement	Adverbial replacement + MI score	
Very interesting	Fascinating, striking, appealing	Extraordinarily interesting	7.80
(MI score 5.97)		Most interesting	6.96
		Particularly interesting	6.41
Very important	Crucial, essential, imperative, serious, vital, remarkable	Vitally important	11.07
(MI score 5.71)		Crucially important	9.14
		Most important	7.91
Very good	Competent, adequate, reliable, kind, real	(Jolly good)	9.51
(MI score 5.79)		(Darn good)	8.60
		(Damn good)	8.45
		Outstandingly good	7.82
		Real good	7.38
		Pretty good	7.18
Very different	Distinct, diverse, unlike	Radically different	9.18
(MI score 5.53)		Markedly different	8.20
		Fundamentally different	7.82
		(Slightly different)	7.33
		Entirely different	6.52
Very happy	Cheerful, glad, thrilled, pleased	Radiantly happy	11.12
(MI score 5.92)		Perfectly happy	7.83
		Quite happy	7.17
Very big	*See separate discussion.*		
(MI score 4.16)			
Very nice	Friendly, kind, lovely	(Jolly nice)	8.55
(MI score 6.71)		Awfully nice	8.09
		Real nice	8.04
Very similar	Akin, identical, related	Uncannily similar	9.41
(MI score 5.56)		Remarkably similar	8.61
		Strikingly similar	8.61
		Startingly similar	8.45
		Somewhat similar	6.74
Very common	Average, ordinary, typical, frequent	Most common	6.80
(MI score 3.81)		Increasingly common	6.34
		Fairly common	6.08
		Quite common	4.85
		(Extremely common)	4.74
Very popular	Famous, trendy, fashionable	Immensely popular	9.35
(MI score 5.60)		Hugely popular	9.23
		Universally popular	8.84
		Wildly popular	6.66
		Highly popular	5.75

6.2.2 Origins of the BT Bigrams

The bigrams that are represented with less than 5 tokens in the BNC1994 have not been assigned a collocational score but are instead grouped as "Below Threshold" (BT). The category deserves a separate discussion, since some of its contents highlight shortcomings of the methods in this

study, and others are of particular interest from a teaching point of view. The examples in the following are all taken from lower secondary school, some from the aggregated material and some from the single student.

Bigrams may fall in the BT category for several reasons, some of which are connected to properties of the chosen reference corpus, the BNC1994. Firstly, while the BNC1994 represents British English, students in Norway are exposed to a massive American cultural influence, and many of them want to speak or write American. Although British English traditionally has had a higher status in Norway, this may be shifting (Rindal, 2010; Kolsvik, 2019). Some expressions used by the students may be more common in American English, because of either linguistic or cultural differences, while they are absent or scarce in the BNC1994 *([Black] Panther party*, *hippie movement*, *dream vacation)*. Secondly, the texts in the BNC1994 are produced prior to or in the early 1990s, making them at least 25 years older than the student texts in the present study. Consequently, some expressions may be absent from the BNC1994 because the thing they refer to did not yet exist at the time of its assembly or had not yet caught the public eye (*social media*, *chat room*, *virtual friends*). Some expressions refer to phenomena that have gained publicity or popularity (*video gaming*, *global warming*, *gay marriage*, *fan fiction*). Thirdly, while the BNC1994 aims to cover a broad spectrum of subjects and genres, it still has its limits, which means expressions may be absent or scarce because they belong to a specialised field which happens to not be represented in the BNC1994 (*menopausal women*, *dear diary*, *shrimp boat*, *Cherokee tribe*). Some expressions may be absent because they are specific for the situation of writing in Norway (*Norwegian parents*, *Sami people*). Fourthly, some expressions may be associated with the situation of writing in a foreign language or in school, making them less likely to appear in the BNC1994 than in a learner corpus (*mock exam*, *preparation material*, *speed presentation*, *global language*). These examples show that the BNC1994 is not an ideal reference corpus for the TRAWL material, and a different reference corpus should be considered for further studies on phraseology in young learner language. Although a subset of the BNC1994 could have been considered a better match for the TRAWL material, this would have left even more bigrams in the BT category, since fewer bigrams would have met the absolute threshold of 5 tokens in the reference corpus.

One expression, *exhilarating taste*, seems to be a novel combination, but made the ≥ 5 range cut because it was provided in the students' preparatory material. Some (near-)absences of bigrams in the BNC1994 appear arbitrary to me. As a non-native speaker of English, I cannot tell whether *ocean view*, *nice city*, and *lunch table* are un-British, unidiomatic, or simply underrepresented.

Most interesting from a pedagogical point of view are probably some expressions provided by writing prompts or preparatory material, that appear non-nativelike. Examples of these are *stereotyped views*, *negative things*, *bad qualities.* These appear to be inaccuracies that the students could have avoided had the writing prompts and preparatory material contained better phrasing. Finally, a number of expressions are absent or scarce in the BNC1994 because they are unidiomatic and represent learner language idiosyncrasies. Some examples, with suggested alternatives in parenthesis, are *dark voice* (*deep voice*), *terror attacks* (*terrorist attacks*), *good laughter* (*good laugh*), *science workers* (*scientists*), *car motor* (*car engine*), *swimming suit* (*swimsuit*).

The material from the single student contains a higher proportion of BT bigrams than the aggregated material, and hints at what can be found if looking beyond the 5-text threshold. This could be worth looking into in future studies focussing on novel combinations or learner idiosyncrasies.

6.2.3 Final Observations

The number of adjective+noun bigrams produced is shown to increase in relative frequency as the learners' level of proficiency increases. Adverb+adjective bigrams associated with spoken language are found in the production at lower proficiency levels, a finding in line with previous research on learner language. Nativelike and non-nativelike uses of words and bigrams are identified, including instances of L1 transfer (*so fun*, *dark voice*, *car motor*). The learners are found to rely to some extent on the material provided for them in assignment wording and preparation texts, especially in 10th grade. This reliance leads to non-native language in cases where the examples provided in the source material are of questionable quality. This illustrates how important it is that the learning material we provide for the students contains high-quality language.

7. Recommendations for Further Research

Controlling for genre and topic of writing is challenging in a longitudinal study of young L2 learners, because general cognitive abilities develop alongside linguistic competence. If feasible, studies controlling for genre and topic could provide great insights into the development of vocabulary and phrase building in young L2 writing.

Only a few word class categories are included in this study. Others could prove interesting, particularly co-occurrences of verb + noun, a category widely studied in the traditional approach to phraseology. Other *n*-gram sizes and discontinuous co-occurring phrases should also be

considered for further study, particularly with a view to identifying items useful for teaching. Other ways of calculating association scores should also be considered, notably MI3 score.

Future statistical approaches to learner phraseology should consider using more suitable reference corpora for defining which expressions are "below threshold", and for assigning association scores to those that are above. The BNC1994, while large and well balanced, is not ideal as a reference corpus for the TRAWL texts. A reference corpus that does not recognise *social media* or *many likes* is not sufficiently up to date to be compared with texts collected in 2017-2019. As changes occur frequently and suddenly in the information society, a monitor corpus should be considered, preferably one where a portion is dedicated to internet and online language. A greater emphasis on American language in the reference corpus is also advisable for comparison to the writing of Norwegian school students.

The present study is corpus-driven and shows, at best, promising patterns in the production data. Although the study provides useful insights for the practice of teaching phraseology, the aggregate method is not sufficient for understanding the development of individual learners. More sophisticated statistical models should be employed to account for individual learner variables and trace their developmental trajectories over time. However, while more sophisticated statistics certainly is desirable, it cannot replace Second Language Acquisition-informed studies of individual learners' development. A study of learner development, aiming to draw conclusions about general developmental trajectories, must be grounded in central aspects of SLA theory.

References

Alley, M. (2018). *The craft of scientific writing* (4th ed.). Springer.

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

Anthony, L. (2019). *AntConc* (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Bestgen, Y., & Granger, S. (2018). Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus. In S. Hoffmann, A. Sand, S. Arndt-Lappe & L. M. Dillmann (Eds.), *Corpora and Lexis* (pp. 277–301). Koninklijke Brill NV.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory, 8*(1), 9–37. https://doi.org/10.1515/cllt-2012-0002

BNC Consortium. (2007). *The British National Corpus, XML Edition*. Oxford Text Archive, http://hdl.handle.net/20.500.12024/2554

Cobb, T., & Horst, M. (2015). Learner Corpora and Lexis. In S. Granger, G. Gilquin & F. Meunier (Eds.). *The Cambridge handbook of learner corpus research* (pp. 185–206). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*(1), 72–89. https://doi.org/10.1093/applin/amm022

Cook, V. (2008). *Second language learning and language teaching* (4th ed.). Hodder.

de Haan, P., & van der Haagen, M. (2013). The search for sophisticated language in advanced EFL writing: A longitudinal study. In S. Granger, G. Gilquin & F. Meunier (Eds.), *Twenty years of learner corpus research: Looking back, moving ahead. Proceedings of the first Learner Corpus Research Conference (LCR 2011)* (pp. 103–115). Presses universitaires de Louvain.

Council of Europe (CEFR) (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press.

Dirdal, H., Hasund, I. K., Drange, E.-M., Vold, E. T., & Berg, E. M. (2022). Design and construction of the Tracking Written Learner Language (TRAWL) corpus: A longitudinal and multilingual young learner corpus. *Nordic Journal of Language Teaching and Learning*, *10*(2), 115–135. https://doi.org/10.46364/njltl.v10i2.1005

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching, 47* (June), 157–177. https://doi.org/10.1515/iral.2009.007

Ebeling, S., & Hasselgård, H. (2015). Learner corpora and phraseology. In S. Granger, G. Gilquin & F. Meunier (Eds.). *The Cambridge handbook of learner corpus research* (pp. 207–229). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. Cambridge University Press. https://doi.org/10.1017/S0272263102002024

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*(March), 17–44. https://doi.org/10.1017/S0267190512000025

Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, *5*(1), 61–78. https://doi.org/10.1515/CLLT.2009.003

Ellis, N. C., Simpson-Vlach, R. & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect 42*(3), 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Ellis, N. C., Simpson-Vlach, R., Römer, U., O'Donnell, M. B., & Wulff, S. (2015). Learner corpora and formulaic language in second language acquisition research. In S. Granger, G. Gilquin & F. Meunier (Eds.). *The Cambridge handbook of learner corpus research*

(pp. 357–378). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Evang, K. H. S. Ø. (2019). *From 'car motor' to 'fishing boat': Tracking intermediate learners' phraseological development by use of association measures.* [Master's thesis, University of Oslo]. http://urn.nb.no/URN:NBN:no-76430.

Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. Oxford University Press.

Gilquin, G., & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction*, *1*(1), 41–61. https://doi.org/10.1075/etc.1.1.05gil

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study." *International Review of Applied Linguistics in Language Teaching*, *52*(3), 229–252. https://doi.org/10.1515/iral-2014-0011

Gries, S. (2015). Statistics for learner corpus research. In S. Granger, G. Gilquin & F. Meunier (Eds.). *The Cambridge handbook of learner corpus research* (pp. 159–181). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Hasselgård, H. (2019). Phraseological teddy bears. In M. Mahlberg, & V. Wiegand (Eds.), *Corpus linguistics, context and culture*. De Gruyter. https://doi.org/10.1515/9783110489071-013

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), 237–259. https://doi.org/10.1111/j.14734192.1994.tb00065.x

Hasund, I. K. (2022). Genres in young learner L2 English writing: A genre typology for the TRAWL (Tracking Written Learner Language) corpus. *Nordic Journal of Language Teaching and Learning*, *10*(2), 242–271. https://doi.org/10.46364/njltl.v10i2.939

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

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *Modern Language Journal* 91(3), 433–445. http://doi.org/10.1111/j.1540-4781.2007.00589.x

Kjellmer, G. (1991). A mint of phrases. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 111–127). Longman.

Kolsvik, S. (2019). *Moving toward(s) Americanization: a study of the use of and attitudes toward American spelling, vocabulary and pronunciation among Norwegian students and teachers.* [Master's thesis, Université catholique de Louvain]. http://hdl.handle.net/2078.1/thesis:18891.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2021). *Introduction to the practice of statistics* (10th ed.). WHFreeman.

Ortega, L. (2013). *Understanding second language acquisition*. Routledge.

Paquot, M. (2013). Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics*, *18*(3), 391–417. https://doi.org/10.1075/ijcl.18.3.06paq

Paquot, M. (2014). Cross-linguistic influence and formulaic language: Recurrent word sequences in French learner writing. In L. Roberts, I. Vedder & J. Hulstijn (Eds.), *EUROSLA Yearbook 14* (pp. 40–261). Benjamins.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In S.C. Richards & R.W. Schmidt (Eds.), *Language and communication* (pp. 191–226). Longman.

Rindal, U. (2010). Constructing identity with L2: Pronunciation and attitudes among Norwegian learners of English. *Journal of Sociolinguistics*, *14*(2), 240–261. https://doi.org/10.1111/j.1467-9841.2010.00442.x

Sinclair, J. (1991). Collocation. In *Corpus, concordance, collocation* (pp. 109–121). Oxford University Press.