Covarage and development of academic vocabulary in English medium instruction

(article starts on next page)

Hans Malmström[1], Diane Pecorari[2], Magnus Gustafsson[1]
[1]Chalmers University of Technology, Gothenburg, Sweden;
[2]Linnaeus University, Växjö, Sweden

Coverage and development of academic vocabulary in assessment texts in English Medium Instruction[1]

Abstract

English: This paper is centred in the context of English Medium Instruction (EMI) and is primarily concerned with advanced students' productive knowledge of English academic vocabulary, widely regarded as a crucial dimension of successful academic communication. The study problematizes the claim that EMI is beneficial for students' development of academic vocabulary knowledge. The investigative context is a technical university in Sweden where all degree programmes at graduate level use English as the medium of instruction. The corpus data include texts (n=80, approx. 720,000 words) produced by Master of Science students in their first and second year of study, written by home and international students. The study, using the Academic Vocabulary List (Gardner/Davis 2014), sets out to answer three research questions relating to knowledge and development of academic vocabulary in EMI: 1. What is the lexical coverage of advanced (master's) level student writing, i.e., what proportion of words in students' texts is academic? 2. Are home students and international students (all of whom have English as a foreign language) comparable in terms of their productive academic vocabulary knowledge? 3. Does students' productive knowledge of academic words appear to develop during their studies? The results of the investigation can be summarized as follows: In the corpus as a whole, academic vocabulary items account for approximately 20% of all tokens. This figure is considerably higher than that found in many earlier studies. There are no significant differences between home and international students in any of the measures of vocabulary used (pertaining to lexical sophistication and diversity). Finally, the findings regarding lexical development across years of study are somewhat mixed; however, the overall picture presented by the various measures is one of significant but very modest gains in some areas and none in others. These

---

findings call into question the actual effectiveness of EMI for academic vocabulary development. The overall contribution of the paper is an important step towards a more comprehensive understanding of what expectations we may reasonably have of the development of English language competency in EMI.

German: Die Fähigkeit, wissenschaftlichen Wortschatz aktiv zu gebrauchen, gilt als eine entscheidende Komponente erfolgreicher Wissenschaftskommunikation. Dabei herrscht die Annahme vor, dass English Medium Instruction (EMI) einen positiven Einfluss auf die Entwicklung des wissenschaftlichen Wortschatzes Studierender habe. Im vorliegenden Beitrag werden die Ergebnisse einer Studie vorgestellt, die diese Annahme kritisch beleuchtet. Die Studie wurde an einer Technischen Universität in Schweden durchgeführt, an der alle Masterstudiengänge auf Englisch unterrichtet werden. Das Korpus umfasst Texte (n=80, ca. 720.000 Wörter), die von schwedischen und internationalen Master-of-Science-Studierenden in ihrem ersten und zweiten Studienjahr verfasst wurden. Die Studie, für die die Academic Vocabulary List (Gardner/Davis 2014) genutzt wurde, geht drei Forschungsfragen nach, in deren Mittelpunkt der Umfang und die Entwicklung wissenschaftlichen Wortschatzes in EMI-Kontexten stehen: 1. In welchem Umfang beherrschen fortgeschrittene Masterstudierende den wissenschaftlichen Wortschatz, d. h., welcher Anteil der Wörter in den studentischen Texten ist wissenschaftlich? 2. Sind schwedische und internationale Studierende mit Englisch als L2 vergleichbar in ihrer Kompetenz, wissenschaftlichen Wortschatz aktiv zu gebrauchen? 3. Entwickelt sich dieser Wortschatz während des Studiums erkennbar weiter? Die Ergebnisse der Untersuchung lassen sich wie folgt zusammenfassen: Im Korpus hat wissenschaftliches Vokabular einen Anteil von ungefähr 20 % aller Tokens und damit einen deutlich höheren Anteil, als in vielen früheren Studien nachgewiesen werden konnte. Im Gebrauch wissenschaftlichen Wortschatzes (sowohl hinsichtlich dessen lexikalischen Anspruchs als auch dessen Differenziertheit) lassen sich zwischen schwedischen und internationalen Studierenden keine signifikanten Unterschiede feststellen. Die Befunde zur Entwicklung des Wortschatzes während des Untersuchungszeitraums sind ambivalent; insgesamt lässt sich jedoch ein moderater Zugewinn in einigen Bereichen feststellen, wohingegen in anderen Bereichen kein Fortschritt zu verzeichnen ist. Diese Ergebnisse geben Anlass zu berechtigten Zweifeln an der Annahme, dass sich mit EMI Wortschatz effizient erweitern lasse. Die Studie leistet somit einen wichtigen Beitrag zur realistischeren Einschätzung der Möglichkeiten, die EMI für die Förderung der englischen Sprachkompetenz bietet.

1 Introduction

The broader context of this study of academic vocabulary knowledge is English Medium Instruction (EMI). For the purposes of this paper, this may be defined as the deliberate use of English (typically as a result of an official educational policy) to engage students communicatively in academic study, i.e., by asking students whose first language is not English to read, write, speak and listen in English rather than using their first language (cf. Coleman 2006; Dearden 2014). While in some contexts a distinction between EMI and Content and Language Integrated Learning (CLIL) is becoming increasingly difficult to maintain, EMI and CLIL should not be considered the same thing. The difference might be one of degree but the defining factor distinguishing the two is the extent to which students receive deliberate language education (CLIL) as opposed to mere immersion (EMI) (Marsh 2005; Lasagabaster 2008; Gustafsson et al. 2011; Gustafsson/Jacobs 2013).

EMI is a "rapidly growing global phenomenon" (Dearden 2014: 2) and a number of different "drivers of the Englishization" (Coleman 2006: 4) of higher education exist. In this respect, there is a widespread assumption articulated to differing extents by the various stakeholders that students enhance their academic as well as general English competency as a result of studying in EMI contexts. However, this is only an assumption (like many other assumptions about EMI reviewed by Dearden), and it has yet to be confirmed by empirical research. Dearden (2014: 2) highlights the fact that "we are quite some way from a 'global' understanding" of EMI and notes that there is an "urgent need for a research-driven approach […] which measures the complex processes involved in EMI", for example, the conditions for "the acquisition of English proficiency". In other words, research that confirms, refutes, or at least problematizes the claim that EMI is beneficial for students' development of English language skills is called for.

In this paper, we are concerned with a single but crucial dimension of English proficiency development in EMI contexts, namely academic vocabulary knowledge, and, more specifically, students' productive knowledge of academic words as reflected in their writing. Our starting point is the widely held claim (see, e.g., Stæhr 2008) that there is a correlation between knowing many words, i.e., having good vocabulary knowledge, and overall communicative competence. Milton (2010: 212) notes that "vocabulary

knowledge is key to both comprehension and communicative ability", and Laufer & Nation (1999: 34) talk of the "enabling" function of vocabulary vis-à-vis other dimensions of communication. In academic discourse, the same kind of correlation obtains between academic words and general academic literacy (see, e.g., Corson 1997; Coxhead 2000; Milton 2010). These correlations are supported by research indicating that understanding virtually all the input words is fundamental to comprehension in any kind of communicative situation (Coady/Huckin 1997; Schmitt 2000; Nation 2001; Bogaards/Laufer 2004). It has been suggested that in excess of 95% of the running words must be understood for "adequate comprehension" to be possible in connection with reading and listening, and for optimal comprehension as much as 98% of the words should be known (Nation 2001; 2006). It seems evident that, if a large vocabulary is needed for reading comprehension, it must be of at least equal importance for the productive assessment tasks in an EMI environment.

In this respect, it is reasonable to ask what might actually be expected of students in EMI in terms of academic vocabulary knowledge. Numerous methods of measuring vocabulary knowledge exist, including self-assessment scales and definition tasks among other measures of receptive knowledge. However, in most EMI settings, students need not only to *understand* English in lectures, textbooks, etc., but also to *produce* it in assessment tasks, so their productive vocabulary knowledge is of interest. Our first research question is therefore:

1. What is the lexical coverage of academic vocabulary in student writing, i.e., what proportion of words in students' texts is academic?

A second perspective we want to explore concerning students' productive knowledge of academic words relates to the Englishization of higher education as a result of globalization/student mobility. One effect of globalization and the concomitant proliferation of EMI has been a rise in EMI outside of the traditionally English-speaking world, in the "Expanding Circle" (Kachru 1992).[2] In Sweden, where this study is set, as in many other

---

2   Kachru (1992) identifies three English "Circles". The Inner Circle is represented by countries like the United Kingdom, USA, Canada, Australia, and New Zeeland, i.e., countries where English is the primary language and the native language of most people. In Outer Circle countries (such as Singapore, India,

Expanding Circle countries, a result has been that significant numbers of 'international' students (students domiciled outside Sweden) are now studying alongside 'home' students (domiciled in Sweden with an almost 100% Swedish language background and English as their first foreign language at school), creating an international mix and a multilingual learning environment in what used to be a linguistically relatively homogenous teaching/ learning environment. The (English) language demands placed on international students and home students in the EMI classroom are naturally the same, but our overall knowledge of how this more diverse group of students actually performs linguistically vis-à-vis home students is very limited.

It is easy to problematize international students and find isolated and categorical statements that speak in general negative terms about international students' shortcomings with regard to English proficiency. For example, in response to a survey about attitudes towards English in higher education administered to university teachers in Sweden (reported in Pecorari et al. 2011), one teacher offered the following comment:

> "English texts, especially academic English texts that we use, are demanding for students. The same is true for writing in English. The problem is especially pronounced for our foreign students who are particularly challenged to write acceptable English."

However, with very few exceptions (see, e.g., Jochems et al. 1996), the research available concerning the overall academic performance and linguistic ability of international students in relation to home students is restricted to Inner Circle countries (see, e.g., Warwick 2006; Carroll/Ryan 2005; Morrison et al. 2005), meaning that our knowledge of how international students compare to home students in Expanding- and Outer Circle countries is almost non-existent.

A first step towards a more comprehensive understanding of the relationship between home students and international students as regards their

---

Pakistan, Bangladesh, Nigeria, and Kenya) English is not the native language but firmly established as the *lingua franca* in most areas of society and typically has the status of "official language". Finally, in countries in the Expanding Circle (the Nordic countries are a case in point), although often widely used for the purpose of international communication (e.g., in much business communication), English is not an official language or the language used in government.

English proficiency is to establish whether home students and international students in EMI environments have equally sized academic English vocabularies, clearly a pertinent investigation given the centrality of academic words to comprehension. This leads us to pose the second of our three research questions:

2. Are home students and international students (all L2 users of English) comparable in terms of their productive vocabulary knowledge?

Finally, this paper directly addresses the supposition that EMI is conducive to developing students' English language skills, focusing on vocabulary gains. There is support in the literature that incidental exposure to English vocabulary in study contexts does lead to positive lexical gains over time (Huckin/Coady 1997; Laufer/Hulstijn 2001); however, the vast majority of this research has focused on learners' receptive knowledge of vocabulary and there is a dearth of research concerned with students' productive lexical knowledge (cf. Durrant 2014). In addition, the bulk of the literature on vocabulary acquisition, and indeed second-language acquisition generally, has focused on learners who are much less advanced than those who are in a position to undertake study at university through the medium of their L2.

Laufer (1994), the most widely cited study on productive vocabulary development available, looked at "changes in the productive lexicon of advanced second language learners' writing over a period of one academic year" (1994: 21) using a construct she calls "lexical quality". Focusing on writing compositions produced in a controlled environment, and by drawing on two basic types of analytical measures, a frequency profile and a Type-Token Ratio, Laufer found no significant lexical gains with regard to general high-frequency words, but there were significant gains for words from the University Word List (Xue/Nation 1984), i.e., Laufer's measure used for academic words, and for words of lower frequency (words beyond the 2,000 most common). With respect to lexical variation, as measured by the Type-Token Ratio, no significant longitudinal gains were recorded for any type of lexis.

Additional longitudinal perspectives on the development of productive lexical knowledge have been provided by more recent research from Australia. This research comes to a different conclusion. Knoch et al. (2015) investigated to what extent the writing of English L2 students developed

positively over three years. Development was measured by looking at a set of discourse-analytic measures, among which was lexical complexity, operationalized with reference to the proportion of words from the Academic Word List (Coxhead 2000), lexical sophistication and lexical richness. For all three measures of lexical complexity, the differences between the first and the second writing collection point fell well short of statistical significance, suggesting that there is little support for the notion that studying in an EMI context has a significantly positive effect on students' productive knowledge of academic lexis.

The setting provided by Laufer's study as well as the study reported by Knoch et al. is in many ways different from EMI contexts in Europe outside Great Britain and Ireland and elsewhere today. It is noteworthy, for example, that the participants in Laufer's investigation were all language learners and therefore possibly 'primed' to attend to linguistic matters like vocabulary, unlike the vast majority of EMI students enrolled in subject courses or degree programmes where there is little or no attention devoted to language *per se*. In addition, it seems unfair to compare the EMI situation of English L2 students studying in English Inner Circle countries with the situation in Expanding- or Outer Circle countries; the complete immersion in an Inner Circle environment presumably affords many more opportunities for engagement with English vocabulary (academic or otherwise). The issue of lexical development in the EMI context facing a great number of students in Expanding- or Outer Circle countries must therefore be investigated independently of such research in Inner Circle countries. Thus, the third research question that this study asks is:

3. Does students' productive knowledge of academic words appear to develop during their studies in an EMI context?

2 Data collection and methods

This section describes the context of the investigation, the data collection and the analytic procedures adopted.

2.1 Study context

Our study is set at a prestigious technical university in Sweden where policy stipulates that all degree programmes at master's level use English as the medium of instruction. The university has approximately 11,000 students, 2,400 of whom are enrolled in one of 41 master's programmes. While we have no reliable record of the first language of the students, the majority of the home students are Swedish L1 speakers and, as far as we have been able to ascertain, none of the international students' whose data were included were domiciled in an English L1 country.

2.2 Data

The data for this study is a small student text corpus made up of 80 texts in English (totalling just over 720,000 running words) written by Master of Science (MSc) students from four different disciplines (applied physics, chemical engineering, chemistry, and mechanical engineering) and from the first and second year of study at the master's level. A total of 30 texts, comprising approximately 115,000 running words, were primarily technical- or mini-project reports written as part of students' course work at some point during the first year. The 50 second-year texts, comprising approximately 605,000 words, were full-length master's theses written during the last term of a two-year study programme. At this university, master's theses are generally reports of project work. Thus, although there is greater variation in the first-year corpus in terms of the assignment set, the two sub-corpora can be regarded as broadly similar in terms of text type.

Because virtually all written course work at this university is done in groups of two or more students, we were unable to obtain first-year and second-year texts from the same author or team of authors. Therefore, 'development' of academic vocabulary refers to change across levels of study rather than change in individual students. In all cases, all authors were either Swedish or international; texts with a mixed authorship with regard to national origin were excluded from the sample in order to enable the comparison between Swedish and international MSc students regarding their productive knowledge of English academic vocabulary.

2.3 Analytical procedure

To address the three research questions, we used several measurements, following Milton (2009) in distinguishing between two basic constructs: lexical sophistication and lexical diversity. The former refers to the extent to which more or less common words are used: in Milton's example, the difference between *the cat sat on the mat* and *the feline reposed on the antique Persian rug* (2009: 131). The latter refers to the extent to which the same or different words are used.

   One measure of lexical sophistication is the presence of academic vocabulary. Over the years, several descriptions and compilations of academic vocabulary have been developed (see Gardner/Davies 2014; and Charles/Pecorari 2016: 109 ff. for a discussion). Until recently, the most widely used list, for both teaching and research, has been Coxhead's (2000) Academic Word List (AWL). As a result of this wide use, a number of limitations of the AWL have been identified,[3] and these limitations have provided the impetus for the newer Academic Vocabulary List (Gardner/Davies 2014); this list is adopted as a basis for what counts as academic vocabulary in the present study.[4]

   The AVL was developed from a 120-million-word academic sub-corpus (featuring texts with a heavy emphasis on journal articles from across nine different academic disciplines) taken from the 425-million-word Corpus of Contemporary American English (COCA 2015). Rather than using an existing word list to exclude general vocabulary (the way Coxhead used the General Service List by West 1953), a criterion for relative frequencies was developed, such that words were considered to be part of an academic core if they occurred in the academic corpus with a frequency 50% greater than in the general, non-academic reference corpus (the non-academic portion of COCA). Words needed to be represented at or above a threshold

---

3   It is not within the scope of this article to criticize the AWL, but Hyland & Tse (2007) and Gardner & Davies (2014) both offer a comprehensive account of the perceived problems with the AWL.

4   The pedagogic utility of lists of general academic vocabulary is widely accepted and Gardner & Davies (2014: 2) note several areas in which such lists are purposeful (see also Schmitt/Schmitt 2014). It should be stressed, however, that pedagogic utility is not a central concern in this study (though see Section 6 where various didactic implications are discussed).

frequency in seven of the nine subject areas, and both a criterion for dispersion and relative frequencies were implemented to exclude words which have a particular affinity with one or a few subject areas (thus allowing for a distinction to be made between core/general academic vocabulary and subject-specific/technical vocabulary). As a result of this process, the AVL consists of 3,015 words (lemmas) that occur across a wide range of academic disciplinary areas more frequently than they do in general discourse. Table 1 includes examples of words from the AVL.

Table 1: Most and least frequent words (lemmas) in the Academic Vocabulary List

AVL words 1–10		AVL words 3006–3015	
study n.	however adv.	unusable adj.	imprimatur n.
group n.	research n.	unpalatable. adj.	coherently adv.
system n.	level n.	causally adv.	component n.
social adj.	result n.	prioritization n.	tangential adj.
provide v.	include v.	overemphasis n.	relevancy n.

Because the AVL contains words that are more common in academic than general discourse, they are in that sense 'advanced' vocabulary. Two measures of lexical sophistication used in the present study were based on the AVL: the proportion of coverage afforded by the AVL, and the number of types from the AVL. However, AVL items vary greatly in frequency, and so not all are equally 'advanced'. For example, the most frequent word on the AVL is *study*, and it also is among the first 1,000 general words by frequency. We thus distinguish between the first 500 words on the AVL and the rest of the list.

The second construct we were interested in was lexical diversity. One of the oldest (Johnson 1939; Mann 1944) and most common measures of lexical diversity is the Type-Token Ratio (TTR). However, the TTR has limitations (Malvern/Richards 2013; Vermeer 2000), including the fact that it is sensitive to text length (Holmes 1994; Baker 2006). While there are no entirely unproblematic measures of lexical diversity, the Guiraud Index (Guiraud 1954) and the Advanced Guiraud (Daller et al. 2003) compensate for the TTR's sensitivity to length and perform more reliably, and were thus adopted here. The former is calculated by dividing the number of types in a text by the square

root of the number of tokens, and the latter uses the same calculation after very common words have been excluded (in this case, the first thousand most frequent words in the BNC and COCA corpora, as provided by Paul Nation's Range files (Range 2015)). Because the advanced measure eliminates the most frequent types, there is a basis for considering it to be an indicator of lexical sophistication as well as diversity, as Daller & Xue (2009) do. Because there is no established baseline for these measures in texts of the type analysed here, they are primarily of value in this study in the two between-group comparisons.

Once the data had been collected, the texts were cleaned[5], converted into text files and processed using AntWordProfiler (Anthony 2015) to determine the frequency of AVL words. In vocabulary profiling (see, e.g., Laufer 1994; Nation 2006) the number of words (tokens) in the texts is counted and the words' distribution relative to pre-established lists is calculated. In this case we used two lists: one list consisting of the 500 most frequent types in the AVL (called AVL 500 here), and one comprising the remaining less frequent lemmas from the AVL (AVL 501+).

Two additional procedures were needed to enable a comparison of the present findings with the Gardner & Davies (2014) study (to the best of our knowledge, the only study based on the AVL to date). Unlike the COCA academic sub-corpus used in that study, our corpus is untagged, meaning that it does not distinguish between words like *study*, n., which is on the AVL, and *study*, v., which is not. To estimate the effect of this difference, a manual search was done among the first 300 words of the AVL for candidates for overcounting (such as *study,* v.). A second procedural issue is that the profiling tools used in the study may have idiosyncrasies which cause them to perform somewhat differently. To estimate the extent of this effect, samples of each corpus text were submitted individually to the lexical profiling tool on Mark Davies' Word and Phrase website (Davies 2015) which analyses the first 1,000 words from each text.

---

5   The following features were removed from the texts: extensive visual information in the form of tables and figures (table and figure captions were left in); all equations/formulae and/or parts thereof, unless some element featured as a syntactic constituent in which case it was treated as technical vocabulary; finally, all tables of contents, reference sections and acknowledgement sections were also removed.

Finally, where relevant, SPSS was used to test the significance of between-group differences. Because a random distribution could not be assumed, a non-parametric test was appropriate (Turner 2014). The independent samples Kruskal-Wallis test was used, and differences were considered significant when $p < .05$.

3 Results

The research questions guiding this investigation related to 1) academic vocabulary coverage; 2) comparisons between home and international students; and 3) comparisons between first-year and second-year texts.

The results of the AVL profiling (Fig. 1), relating to the lexical sophistication of the texts, showed that 19.3% of the tokens in the corpus are academic words. This is a considerably higher proportion of academic vocabulary than previous studies have shown. The most relevant earlier study is Gardner & Davies (2014), who found that the AVL gave coverage of the academic sections of the COCA and BNC in the vicinity of 14%. To estimate the extent to which procedures may have contributed to these different results, two additional analyses were conducted.
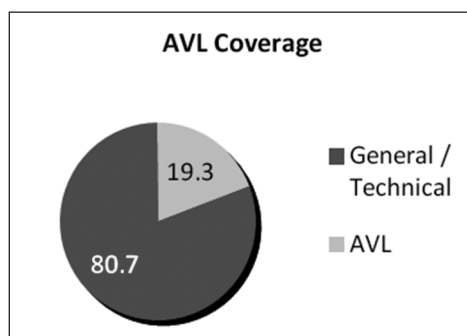

Fig. 1: Academic vocabulary coverage in the MSc writing corpus

To account for the effect of the untagged corpus, a manual search was done among the first 300 words of the AVL for candidates for overcounting (such as *study,* v.). A total of 2,028 such forms were identified, or approximately 1.4% of the total corpus size. The effect of the untagged corpus is therefore real but relatively modest. When the first 1,000 words of each text were

submitted to the Word and Phrase profiling tool (Davies 2015), an average of 23% of the tokens came from the AVL. It is therefore reasonable to conclude that academic words do in fact make up approximately 20% of our corpus and that procedural issues play a relatively small role. Other explanations for the difference between these findings and earlier studies are taken up in Section 4 below.

A further measure of lexical sophistication was the proportion of infrequent AVL words. In the COCA academic corpus, the 500 most frequent of the 3,015 AVL types (i.e., 17%) account for 74% of all of AVL tokens. The average number of tokens representing each type in the 1–500 list was 14 times greater than on the 501+ list (22,599 versus 1,480) (see Fig. 2). As Figure 2 shows, the figures for the present corpus are comparable: 70.5% of the AVL tokens come from the first 500 words of the AVL, with only 29.5% coming from the remainder of the list, and the average type on the 1–500 list had nearly 13 times as many tokens as the average type on the 501+ list.
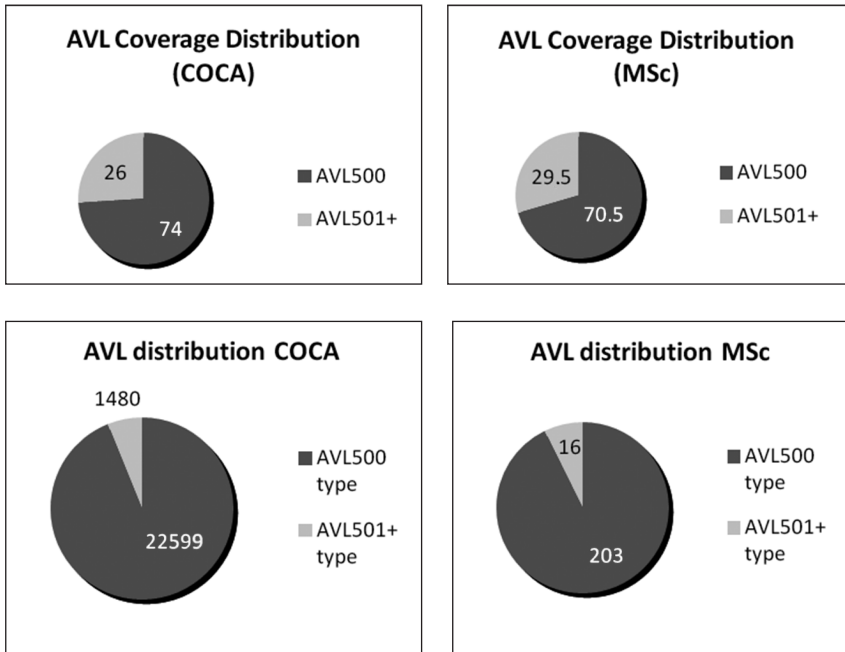

Fig. 2: *Academic vocabulary diversity and sophistication in the COCA and MSc writing corpora*

The second research question addressed the relative productive vocabulary of native Swedish and international students. Given that large numbers of international students are a relatively recent phenomenon in Swedish university classrooms, there is a need to understand whether the English skills of this new constituency permit them to participate in EMI on the same terms as their local counterparts. Measures for lexical sophistication and lexical diversity were therefore considered for the two groups separately.

As Table 2 shows, for all measures, the differences were small, and indeed none of them was statistically significant. This measure thus suggests, encouragingly, that this relatively new student group is able to take on EMI education on a level playing field with their Swedish peers, at least when assessed on the basis of the productive vocabulary knowledge.

Table 2: AVL distribution for home versus international students

Measure	Overall	Home	International
AVL coverage (entire list)	19.3%	18.9%	19.9%
AVL 500 coverage	13.6%	13.2%	14.1%
AVL 501+ coverage	5.74%	5.7%	5.8%
AVL Types per 100 words	5.69	5.88	5.40
AVL 500 types per 100 words	3.60	3.76	3.37
AVL 501+ types per 100 words	2.08	2.12	2.03
Guiraud Index	14.2	14.01	14.48
Advanced Guiraud	8.68	8.33	9.20

Since language development is one of the reasons offered for implementing EMI, it would be reasonable to think that students' vocabulary – particularly academic vocabulary – develops during their course of study. The third research question was therefore whether the second-year texts showed greater lexical sophistication and variation than the first-year texts.

The measures of lexical sophistication failed to reflect gains between the two groups. As Table 3 shows, a small increase was found for the overall coverage afforded by the AVL, from 19.0% to 19.5%. On closer investigation, this increase is seen to be driven by increased usage of the 500 most frequent AVL items, from 12.9% to 13.9%, which was in turn offset by a

slight decrease in coverage from the remainder of the list. However, none of these changes were significant.

The normalized frequencies of types from the entire AVL, the most frequent 500 items and the less frequent items, all showed a significant decrease in the second year (p=.000). This (in combination with the change in the Guiraud Index and Advanced Guiraud, see below) is an indication that the greater diversity of lexis came either from general (i.e., non-academic) vocabulary in the less frequent range, and/or from technical terminology, rather than increased usage of general academic vocabulary.

In terms of variation, both the Guiraud Index and the Advanced Guiraud showed a modest but significant (p=.000) trend toward greater lexical variation in the year-two texts (see Table 3). Because it excludes the first thousand most commonly used words, the Advanced Guiraud also reflects lexical sophistication.

Table 3: AVL distribution in first- versus second-year texts

Measure	Overall	Year 1	Year 2
AVL coverage (entire list)	19.3%	19.0%	19.5%
AVL 500 coverage	13.6%	12.9%	13.9%
AVL 501+ coverage	5.74%	6.0%	5.6%
AVL Types per 100 words	5.69	7.40	4.73
AVL 500 types per 100 words	3.60	4.85	2.90
AVL 501+ types per 100 words	2.08	2.55	1.82
Guiraud Index	14.2	12.96	14.90
Advanced Guiraud	8.68	7.46	9.37

4 Discussion

Section 3 presented findings that were in some ways unexpected, and thus merit further discussion.

4.1 Academic vocabulary coverage

The finding that approximately 20% of this corpus consisted of academic words contrasts strikingly with the much lower figures in previous studies. For example, Coxhead (2000) and Hyland & Tse (2007) found that

the Academic Word List afforded about 10% coverage of their respective corpora. A 20%-figure was approached only in Chung & Nation's (2003) study concerning Applied Linguistics textbooks. As noted above, methodology accounts for only a minor part of the difference, and it is interesting to speculate as to what may account for the rest.

One likely explanation is the composition of the corpus; the proportion of academic vocabulary varies according to text type (Chung/Nation 2003; Li/Qian 2010) and academic discipline (Chung/Nation 2003; Coxhead 2000; Hyland/Tse 2007). However, no other study based on a fully comparable corpus exists. Engineering was one of the fields investigated by Hyland & Tse (2007) and Mudraya (2006), but the former corpus contained a mix of text types, while the latter consisted of textbooks and provided no overall academic vocabulary coverage figure. Thus, while it is probable that academic subject area and text type explain some of the difference between the present findings and earlier ones, it is not possible to ascertain the extent of their influence.

A second explanation lies in the use here of the Gardner & Davies (2014) AVL, while previous investigations have employed Coxhead's (2000) AWL. The AVL's ability to represent core academic vocabulary better than the AWL has been demonstrated empirically. Gardner & Davies (2014) profiled the academic sections of the BNC and COCA with both the AWL and the AVL, using word families to enable a comparison with the AWL, and found that the top AVL 570 word families provided nearly twice the coverage with respect to the AWL (13.8% versus 7.2% in COCA; 13.7% versus 6.9% in the BNC). In this light, the fact that the present study found approximately twice as much academic vocabulary as earlier studies is unsurprising; indeed an investigation of the present MSc writing corpus (with some modifications) found that the AWL provided just under 10% coverage (Gustafsson/Malmström 2013).

A further question is why AVL coverage is higher for the MSc writing corpus than the academic portions of COCA and the BNC. Here too, corpus composition undoubtedly plays a role. In addition, there is likely to be an effect due to an aspect of Gardner & Davies' (2014) methodology. While the AVL (unlike the AWL) is not based on word families, their figure of approximately 14% coverage comes from a case study which, in order to permit comparisons with the AWL, used part of the AVL grouped

into 570 word families. The findings of these studies are therefore not fully comparable, and because of the lack to date of studies using the AVL, further research is needed.

4.2 Home versus international students

There is a fairly commonplace belief on the part of many Swedish university teachers that international students have lower English proficiency than home students. This perception is frequently offered almost apologetically; there is a widespread perception that international students enrich the Swedish university classroom and that their presence is therefore desirable, but that achieving an international student presence requires the use of English as an academic *lingua franca*, and while this puts all participants at a disadvantage, those who have gone through the Swedish educational system, which emphasizes English, are better able to cope than most incoming mobile international students. It is not clear how to explain the disparity between this belief and the findings of the present study.

A possible explanation is that teacher perceptions are based less on reality and more on an awareness of differences. More specifically, the English used by Swedish university students is familiar to their teachers, and the non-standard transfer features that characterize it are unmarked, while those of students with other origins are more salient. Another possibility, which indeed is applicable to all of the findings reported here, is that students recruited to the prestigious university where this study was conducted are a relatively homogeneous, skilled group of English users. Were the study to be replicated at another institution, between-group differences might be identified. It is also possible that these groups may differ in English proficiency, but that the differences manifest themselves in other domains than productive academic vocabulary (i.e., in other domains of oral and written communication). Future research would be required to establish the extent to which any of these explanations is a factor.

4.3 Vocabulary development over time

One of the intended benefits of EMI is that it creates exposure to the language and can therefore result in incidental vocabulary acquisition. Academic vocabulary would appear to be a prime candidate for such ac-

quisition, since it is an area of language to which students can be expected to have greatest exposure in a university setting. It is therefore somewhat counterintuitive that the findings for academic vocabulary development were mixed.

One reason for this may be that even the least experienced writers in this study were highly proficient. By virtue of being deemed capable of doing postgraduate academic work through the medium of English, these students can be classed as advanced users of English, and this is additionally indicated by the fact that their texts were richly populated with academic vocabulary. As Hyltenstam & Abrahamsson (2012) note, research on very proficient L2 learners is in short supply compared with the voluminous body of second language acquisition research on learners at lower proficiency levels. However, it is reasonable to expect their learning to progress at a slower pace, simply because they have less ground to cover. In other words, there may be a phenomenon at play akin to a ceiling effect, according to which the year two texts did not show much greater lexical diversity and sophistication because the year one texts were already satisfactory in that regard.

Similarly, it may be thought that these students had relatively limited opportunities for vocabulary development. The EMI environment provides a context in which only incidental language acquisition can occur, rather than an EAP/TEFL environment where language development is the target of explicit instruction. As a result, opportunities for language learning are closely linked to *exposure* to the linguistic features that are candidates for learning. Less proficient learners have more opportunities for exposure to new forms than advanced learners, precisely because more of what they are exposed to is new. In the case of the high-register academic vocabulary that was the focus of the present investigation, the opportunities for exposure to the infrequent words decrease logarithmically, not arithmetically, once the first bands of very frequent words have been learned.

5 Conclusions

This article has reported an investigation into the academic vocabulary knowledge of students in an EMI setting. Students' knowledge of academic vocabulary is important in this context because it is essential both

for adequate comprehension of academic texts and for producing register-appropriate assessment work. As a consequence of the fact that study in the EMI environment places demands on students' receptive and productive academic vocabulary knowledge, it is an aspect of linguistic proficiency which could reasonably be expected to develop over the course of their studies. A measure of students' productive academic vocabulary is therefore a useful indicator (though by no means the only one) of two important factors: students' preparedness for academic study, and their development in English.

To the extent that the findings presented in this paper speak to preparedness, they permit an optimistic interpretation: academic vocabulary items accounted for approximately 20% of all tokens, a rather higher figure than that found in many earlier studies. Although knowledge of academic vocabulary alone cannot be interpreted as evidence that students are equal to the challenges of study through the medium of English, a more cautious claim can be made: there is no reason to believe that this cohort of students lacks an adequate knowledge of academic vocabulary.

The high level of coverage also provides support for the principles underlying the construction of the AVL. By including items which occur in academic texts more frequently than in general ones, and by excluding items which occur disproportionately frequently in some disciplines only, the AVL is designed to give a better representation of general academic vocabulary than earlier lists, and the incidence of AVL items in the present corpus provides indirect evidence that the AVL behaves the way it was intended. While this does not resolve all of the problematic aspects of the notion of an academic core vocabulary (cf. Hyland/Tse 2007), it suggests that, in circumstances where an academic vocabulary list is necessary or desirable, for pedagogical or research purposes, the AVL is the list of choice.

Perhaps more significantly, this measure of productive academic vocabulary gives no support for the idea that international students and local Swedish students differ in their abilities in English. This is reassuring given the fact that the economic and policy imperatives in Swedish higher education (and reflected elsewhere in Europe) will for the foreseeable future lead to an increase in inward student mobility.

With regard to vocabulary development between the first and the second year, evidence was limited; there were modest gains by some measures

and none by others. This is a finding of relevance given the current rapid expansion of EMI, and the twin motivations behind it. EMI is expected to be both a tool to facilitate mobility in higher education and a vehicle for improved English language skills on the part of participants but this study of academic vocabulary knowledge provides little indication that the latter ambition is realized, at least in the context under investigation.

6 Pedagogical implications

In this volume, with its focus on the pedagogical aspects of assessment, the pedagogical implications of the findings merit exploration. However, the EMI environment is complex in its pedagogical objectives. One objective of EMI is simply to enable the teaching and learning of subject matter by using English as an academic *lingua franca*. In many EMI contexts, though, an additional objective is to provide a context which facilitates students' incidental acquisition of English. The pedagogical implications of students' vocabulary knowledge and development are different for these two different objectives.

In terms of content learning, these students appear to be well equipped with a productive knowledge of academic vocabulary sufficient to complete assessment tasks (and therefore by implication with a receptive vocabulary sufficient to read academic texts). This means that teachers (provided they have similar student profiles and communication genres) can concentrate on, for instance, promoting the critical reading of the disciplinary vocabulary. From a collaborative learning perspective, peer learning can enable the further exploration of the enhanced understanding of technical vocabulary.

In the scenarios where the EMI context involves an element of collaboration or contact between language lecturers and subject lecturers, the language lecturer might help the subject lecturer highlight the way in which academic vocabulary serves to carry the disciplinary argument. Such a shared focus would help students articulate the necessary disciplinary connections between argumentative components. A subject lecturer might contribute with useful insights for prompts, exercises, and classroom assessment techniques focused on exploring technical vocabulary.

With respect to language development, teachers may conclude that basic academic vocabulary knowledge can be taken as confirmed. They can there-

fore use this apparent communicative resource of academic vocabulary as a stepping-stone to explore the remaining dimension of written disciplinary communication. For example, they might have students extract and master the technical vocabulary in the texts they encounter via basic critical reading using genre and corpus analyses.

The productive knowledge of the frequent academic vocabulary items demonstrated here could also be a potential stepping-stone toward command of the less frequent AVL items. However, the evidence of this study is that development along those lines does not happen automatically, and indeed there is no reason to suppose it should, given that opportunities for exposure to infrequent vocabulary are limited. A key pedagogical implication of these findings is therefore that *incidental* acquisition is unlikely to be *accidental*, and that teachers who hope their students' academic vocabulary will develop during an EMI course should create opportunities for exposure to and practice of a broader range of academic lexis.

This study underscores a reality of many EMI settings. EMI is intended to be a *de facto* form of Content and Language Integrated Learning (CLIL) but while CLIL settings work actively both with content knowledge and with language development, in EMI the expectation is frequently that the preconditions for incidental language acquisition are put in place simply by dint of offering instruction in English. This study has provided evidence that those expectations are not entirely justified.

References

Anthony, Laurence, 2015: *AntWordProfiler (Version 1.4.1)* [Computer Software]. Tokyo, Japan: Waseda University. http://www.laurenceanthony.net/ (last accessed: 30 July, 2015).

Baker, Paul, 2006: *Using Corpora in Discourse Analysis*. London: Continuum.

Bogaards, Paul / Laufer, Batia (eds), 2004: *Vocabulary in a Second Language: Selection, Acquisition, and Testing* (Vol. 10). Amsterdam: John Benjamins Publishing.

Carroll, Judith / Ryan, Janette (eds), 2005: *Teaching International Students – Improving Learning for all*. Abingdon: Routledge.

Charles, Maggie / Pecorari, Diane, 2016: *Introducing English for Academic Purposes*. London: Routledge.

Chung, Teresa Mihwa, / Nation, Paul, 2003: "Technical vocabulary in specialised texts." *Reading in a Foreign Language* 15(2), 103–116.

Huckin, Thomas / Coady, James, 1997: *Second Language Vocabulary Acquisition: A Rationale for Pedagogy*. Cambridge: Cambridge University Press.

COCA, 2015: *Corpus of Contemporary American English*. [Computer software]. http://corpus.byu.edu/coca (last accessed: 30 July, 2015).

Coleman, James A., 2006: "English-medium teaching in European higher education." *Language Teaching* 39(1), 1–14.

Corson, David, 1997: "The learning and use of academic English words." *Language Learning* 47(4), 671–718.

Coxhead, Averil, 2000: "A new academic word list." *TESOL Quarterly* 34(2), 213–238.

Daller, Helmut / van Hout, Roeland / Treffers-Daller, Jeanine, 2003: "Lexical richness in the spontaneous speech of bilinguals." *Applied Linguistics* 24(2), 197–222.

Daller, Helmut / Xue, Huijuan, 2009: "Vocabulary knowledge and academic success: A study of Chinese students in UK higher education." In: Richards, Brian / Daller, H. Michael / Malvern, David D. / Meara, Paul / Milton, James / Treffers-Daller, Jeanine (eds): *Vocabulary Studies in First and Second Language Acquisition: The Interface Between Theory and Applications*. Basingstoke: Palgrave Macmillan, 179–193.

Davies, Mark, 2015: *Word and Phrase*. [Computer software]. http://www.wordandphrase.info/academic/ (last accessed: 30 July, 2015).

Dearden, Julie, 2014: *English as a Medium of Instruction – a Growing Global Phenomenon*. (Research report). http://www.education.ox.ac.uk/wordpress/wp-content/uploads/2014/09/EMI-a-Growing-Global-Phenomenon-new-cover.pdf (last accessed: 30 July, 2015).

Durrant, Phil, 2014: "Discipline and level specificity in university students' written vocabulary." *Applied Linguistics* 35(3), 328–356.

Gardner, Dee / Davies, Mark, 2014: "A new academic vocabulary list." *Applied Linguistics* 35(3), 305–327.

Guiraud, Pierre, 1954: *Les caracteres statistiques du vocabulaire*. Paris: Presses Universitaires de France.

Gustafsson, Magnus / Eriksson, Andreas / Räisänen, Christine / Stenberg, Ann-Charlotte / Jacobs, Cecilia / Wright, Jenny / Wyrley-Birch, Bridget / Winberg, Chris, 2011: "Collaborating for content and language integrated learning: The situated character of faculty collaboration and student learning." *Across the Disciplines* 8(3). http://wac.colostate.edu/atd/clil/gustafssonetal.cfm (last accessed: 30 July, 2015).

Gustafsson, Magnus / Jacobs, Cecilia, 2013: "Student learning and ICLHE – frameworks and contexts." *Journal of Academic Writing* 3(1). http://e-learning.coventry.ac.uk/ojs/index.php/joaw/article/view/141 (last accessed: 30 July, 2015).

Gustafsson, Magnus / Malmström, Hans, 2013: "Master level writing in engineering and productive vocabulary: What does measuring academic vocabulary levels tell us?" In: Johannesson, Nils-Lennart / Melchers, Gunnel / Björkman, Beyza (eds): *Of Butterflies and Birds, of Dialects and Genres: Essays in Honour of Philip Shaw*. Stockholm: Stockholm Studies in English 104, 123–140.

Holmes, David I., 1994: "Authorship attribution." *Computers and the Humanities* 28(2), 87–106.

Hyland, Ken / Tse, Polly, 2007: "Is there an 'academic vocabulary'?" *TESOL quarterly* 41(2), 235–253.

Hyltenstam, Kenneth / Abrahamsson, Niclas, 2012: "Introduction." *Studies in Second Language Acquisition* 34(2), 177–186.

Jochems, Wim / Snippe, Joke / Smid, Harm Jan / Verweij, Agnes, 1996: "The academic progress of foreign students: study achievement and study behaviour." *Higher Education* 31(3), 325–340.

Johnson, Wendell, 1939: *Language and Speech Hygiene: An Application of General Semantics. Outline of a Course*. General semantics monographs 1. Lakeville, CT: Institute of General Semantics.

Kachru, Braj, B., 1992: *The Other Tongue: English Across Cultures*. Champaign, IL: University of Illinois Press.

Knoch, Ute / Rouhshad, Amir / Oon, Su Ping / Storch, Neomy, 2015: "What happens to ESL students' writing after three years of study at an English medium university?" *Assessing Writing* 28, 39–52.

Lasagabaster, David, 2008: "Foreign language competence in content and language integrated courses." *The Open Applied Linguistics Journal* 1(1), 30–41.

Laufer, Batia, 1994: "The lexical profile of second language writing: does it change over time?" *RELC Journal* 25(2), 21–33.

Laufer, Batia / Hulstijn, Jan, 2001: "Incidental vocabulary acquisition in a second language: The construct of task-induced involvement." *Applied Linguistics* 22(1), 1–26.

Laufer, Batia, / Nation, Paul, 1999: "A vocabulary-size test of controlled productive ability." *Language Testing* 16(1), 33–51.

Li, Yongyan / Qian, David, 2010: "Profiling the Academic Word List (AWL) in a financial corpus." *System* 38(3), 402–411.

Malvern, David / Richards, Brian, 2013: "Measures of lexical richness." In: Chapelle, Carol A. (ed.): *The Encyclopedia of Applied Linguistics*. Oxford, UK: Blackwell Publishing Ltd, 3622–3627.

Mann, Mary Bachmann, 1944: "The quantitative differentiation of samples of written language." *Psychological Monographs* 56(2), 39–74.

Marsh, David, 2005: Project D3 – CLILMatrix – Central workshop report 6/2005 (Graz, 3–5 November 2005). European Centre for Modern Languages 6. http://archive.ecml.at/mtp2/CLILmatrix/pdf/wsrepD3E2005_6.pdf (last accessed: 30 July, 2015).

Milton, James, 2009: *Measuring Second Language Vocabulary Acquisition*. Bristol: Multilingual Matters.

Milton, James, 2010: "The development of vocabulary breadth across the CEFR levels." *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research*, 211–232.

Morrison, Jo / Merrick, Beatrice / Higgs, Samantha / Le Métais, Joanna, 2005: "Researching the performance of international students in the UK." *Studies in Higher Education* 30(3), 327–337.

Mudraya, Olga, 2006: "Engineering English: A lexical frequency instructional model." *English for Specific Purposes* 25(2), 235–256.

Nation, Paul, 2001: *Learning Vocabulary in Another Language*. Cambridge, England: Cambridge University Press.

Nation, Paul, 2006: "How large a vocabulary is needed for reading and listening?" *Canadian Modern Language Review/La Revue Canadienne des Langues Vivantes*, 63(1), 59–82.

Pecorari, Diane / Shaw, Philip / Irvine, Aileen / Malmström, Hans, 2011: "English for Academic Purposes at Swedish universities: teachers objectives and practices." *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE)* 22, 55–78.

Range, 2015: [Computer software]. Available at: http://www.victoria.ac.nz/lals/about/staff/paul-nation.

Schmitt, Norbert, 2000: *Vocabulary in Language Teaching*. Cambridge, England: Cambridge University Press.

Stæhr, Lars Stenius, 2008: "Vocabulary size and the skills of listening, reading and writing." *Language Learning Journal* 36(2), 139–152.

Turner, Jean L., 2014: *Using Statistics in Small-Scale Language Education Research: Focus on Non-Parametric Data*. New York: Routledge.

Vermeer, Anne, 2000: "Coming to grips with lexical richness in spontaneous speech data." *Language Testing* 17(1), 65–83.

Warwick, Philip, 2006: *International Students in the UK: How Can We Give Them a Better Experience?* (Working Paper). Department of Management Studies, University of York, York.

West, Michael, 1953: *A General Service List of English Words*. London: Longman, Green & Co.

Xue, Gua-yi / Nation, Paul, 1984: "A university word list." *Language Learning and Communication* 3(2), 215–29.