Profiling lexical coverage in EMI academic seminars: a corpus-based study

Irene Castellano-Risco
University of Extremadura (Spain)
ircastellano@unex.es

Abstract

Since the implementation of the Bologna Process, Higher Education Institutions have increasingly offered English as the Medium of Instruction (EMI) programmes to promote "internationalisation at home" (Beelen & Jones, 2015). While these programmes allow students to take content courses in English, EMI settings may also be challenging for learners due to the linguistic demands they may impose. However, current efforts to identify these demands have primarily relied on self-reported data, with limited consideration given to the linguistic features that may cause difficulties for EMI learners. This paper aims to address this gap by exploring the lexical demands of EMI seminars, a teaching situation characterised by being student-centred and dialogic in nature. The study analyses the lexical coverage of BNC/COCA (Nation, 2012) and ASW (Dang et al., 2017) lists in METCLIL, a corpus comprising nine EMI seminars and 111,061 tokens, and investigates the impact of contextual variables on lexical demands. The findings indicate that (1) EMI seminars show lower lexical demands on L2 learners than L1 lectures and seminars or EMI lectures but similar to L1 tutorials, (2) academic spoken vocabulary is less frequent than in other educational settings, and (3) certain contextual factors seem to play a critical role in lexical coverage.

Keywords: EMI, academic vocabulary, lexical coverage, higher education, L2 proficiency.

Resumen

Perfilando la cobertura léxica en seminarios impartidos en inglés como medio de instrucción (EMI): un estudio de corpus

Desde la implantación del proceso de Bolonia, los centros de Educación

Superior han venido aumentando su oferta de programas en inglés como medio de instrucción (*English Medium Instruction*, EMI) para fomentar la "internacionalización en casa" (Beelen & Jones, 2015). Si bien este tipo de programas permite a los estudiantes cursar asignaturas en inglés, también pueden suponerles un desafío debido a las exigencias lingüísticas que pueden imponer. Los esfuerzos realizados para identificar estas exigencias se han basado fundamentalmente en las percepciones de los propios participantes y se le ha prestado muy poca atención a las características lingüísticas que pueden generar dificultades a los estudiantes de EMI. Este artículo pretende abordar esta laguna explorando las exigencias léxicas en seminarios EMI, un contexto educativo que se caracteriza por estar centrado en el alumnado y tener una naturaleza dialógica. Este estudio analiza la cobertura léxica de las listas BNC/COCA (Nation, 2012) y ASW (Dang et al., 2017) en METCLIL, un corpus que está compuesto por nueve seminarios EMI, con un total de 111.061 palabras, e investiga el impacto de las variables contextuales en las exigencias lingüísticas. Los resultados indican que (1) los seminarios EMI muestran menores demandas léxicas para los estudiantes de L2 que las clases magistrales y seminarios de L1 o las clases magistrales EMI, pero estas demandas son similares a las tutorías de L1; (2) el vocabulario académico oral es menos frecuente que en otros contextos educativos; y (3) ciertos factores contextuales parecen desempeñar un papel crítico en la cobertura léxica.

Palabras clave: EMI, vocabulario académico, cobertura léxica, educación superior, dominio de la L2.

1. Introduction

"Internationalisation at home" practices (Beelen & Jones, 2015) have gained popularity in recent years, with a focus on promoting international understanding and intercultural competence among students at their home universities (Beelen & Jones, 2015; Knight, 2006). One of the most widely adopted practices within this trend is the implementation of English as a Medium of Instruction (EMI) programmes, which involve teaching courses in English at institutions where it is not the official language.

EMI programmes offer English L2 students the opportunity to engage in content-based lectures, seminars, and lab sessions conducted in English (Macaro, 2018). These programmes provide a platform for active participation and meaningful use of the L2 but also pose a range of linguistic and cognitive demands on learners that are often overlooked. Consequently, extensive research has been conducted to identify the primary linguistic

needs of lecturers and students (Coelho, 2022; Pérez-Cañado, 2021; Piquer-Píriz & Castellano-Risco, 2021), particularly in the areas of classroom language and academic language development (Piquer-Píriz, 2023).

The existing literature suggests that students in EMI programs often face challenges in understanding lectures, which are often attributed to their limited L2 vocabulary (Hellekjaer, 2010; Shepard & Morrison, 2021). Nonetheless, the reported needs of EMI learners are primarily based on self-assessment and lack empirical evidence from research conducted in a classroom setting. To address this issue, researchers can employ lexical studies to verify the self-reported needs of EMI students. Previous studies in second language acquisition have shown that lexical coverage, which refers to the percentage of recognised lexical items in a written or oral text, is a crucial factor in L2 comprehension (Laufer & Sims, 1985; Schmitt et al., 2011). This factor may be particularly relevant in EMI settings, where students need to understand new concepts and ideas, while processing input that is unlikely to be comprehended without a broad L2 vocabulary knowledge.

For this reason, this study aims to contribute to EMI linguistic-oriented research by determining the lexical demands of EMI academic seminars and exploring how contextual factors may affect coverage in this particular learning context. The implications of this research will be useful for EMI learners and educators, as they can help them better understand the specific vocabulary requirements for success in academic seminars and develop more effective vocabulary teaching and learning strategies.

The paper is structured as follows. Section 2 contextualises the study by highlighting the significance of lexical studies for the advancement of EMI and reviewing previous research on vocabulary demands of academic settings. Then, section 3 details the methodological approach followed. It presents METCLIL, a corpus of 9 EMI seminars on the field of marketing and business, and the main instruments employed to analyse the lexical coverage of the corpus. After that, section 4 presents the results arranged according to the three RQs posed in the methodological explanation, and section 5 discusses the main findings reflecting about their pedagogical implications.

2. Lexical coverage and EMI

Vocabulary studies, i.e., the analysis of the role of vocabulary in communication and how individuals acquire it, have become a prominent

field of research in recent decades (c.f., Nation, 2013; Durrant et al., 2022, or Schmitt, 2010), resulting in a reconsideration of the relevance of L2 lexis for effective communication.

L2 learners need to manage a broad range of vocabulary to communicate efficiently (Nation, 2013; Schmitt, 2010), and in their endeavour to guide students, L2 practitioners need to recognise and identify (i) the lexical difficulty of the oral and written texts learners face, and (ii) the vocabulary students need to communicate in a foreign language. In this respect, research has proved that the more extensive the lexical coverage of a text, the larger the comprehension (Schmitt et al., 2011).

The literature (Vilkaite-Lozdiene & Schmitt, 2020; van Zeeland & Schmitt, 2013) suggests the existence of at least two lexical thresholds depending on the coverage aimed at: an "optimal" one (Vilkaite-Lozdiene & Schmitt, 2020), also referred to as "very high comprehension" (van Zeeland & Schmitt, 2013, p. 474), that places the coverage needed at 98% (Hu & Nation, 2000; Nation, 2006; Schmitt et al., 2011; van Zeeland & Schmitt, 2013), and a "minimal" one (Vilkaite-Lozdiene & Schmitt, 2020), also known as "relatively high and stable comprehension" (van Zeeland & Schmitt, 2013, p. 474), in which the percentage of understanding decreases to 95% (Laufer & Ravenhorst-Kalovski, 2010; van Zeeland & Schmitt, 2013; Webb & Rodgers, 2009a, 2009b).

The thresholds for vocabulary recognition help determine the number of lexical items L2 learners need to master. This is typically expressed in word families, a lexical unit consisting of a base form and all its derived and inflected forms (Bauer & Nation, 1993). For everyday communication and general English, it is estimated that learners need to master between the 2,000 to the 3,000 most frequent word families (Adolphs & Schmitt, 2003; Nation, 2006) for 95% coverage. Regarding 98% coverage, there is some debate about the specific number of word-families needed, ranging from the 3,000 most frequent word-families (Adolphs & Schmitt, 2003; Webb & Rodgers, 2009a) to the 6,000 to 7,000 most frequent word families (Hirsh & Nation, 1992; Nation, 2006; Webb & Rodgers, 2009a). Based on these findings, Schmitt and Schmitt (2014) divided word families into three major categories: high-frequency word families (1st to 3,000th word families), mid-frequency word families (3,001st word family to 9,000th) and low-frequency word families. This distinction is mainly for pedagogical purposes, as high-frequency terms provide 95% coverage, while 98% coverage typically requires the recognition of mid-frequency word families.

The concept of vocabulary frequency and its implications for L2 listening and reading soon attracted the attention of fields such as English for Academic Purposes (EAP), in which there have been some efforts to distinguish core academic vocabulary from General English and develop specific word lists, such as the University Word list (UWL; Xue & Nation, 1984), the Academic Word List (AWL; Coxhead, 2000); the Academic Vocabulary List (AVL; Gardner & Davies, 2014) or, more recently, the Academic Spoken Word List (ASWL; Dang et al., 2017). These lists have traditionally served two main purposes: (1) to aid publishers and EAP instructors in creating specific educational materials, and (2) to provide information about the extent of academic vocabulary usage in academic genres.

Concerning this latter purpose, the analysis of general and academic vocabulary coverage has been applied to academic settings in a variety of academic genres, such as English L1 lectures and seminars (Dang & Webb, 2014), English L1 tutorials and laboratories (Coxhead et al., 2017), TED talks (Liu & Chen, 2019; Nurmukhamedov, 2017), conference presentations (Dang, 2022a), or English L1 open-courses and EMI lectures (Dang, 2022b). Table 1 summarises some recent studies on several academic genres.

Study	Genre	BNC/COCA list*		Academic Spoken Word List (ASWL)			
		95%	98%	Level 1	Level 2	Level 3	Level 4
Dang & Webb, 2014	L1 lectures	4,000	8,000	-	-	-	-
Dang et al., 2017	13.5-million-word academic spoken corpus			81.62	5.23	2.85	0.43
Coxhead et al., 2017	Laboratories Tutorials	3,000 2,000	7,000 3,000	- -	- -	- -	- -
Nurmukhamedov, 2017	TED talks	4,000	8,000				
Liu & Chen, 2019	TED talks	3,000	6,000	82.5%	4.7%	2.2%	0.2%
Dang, 2022a	Hard science presentations Soft science presentations	3,000 3,000	5,000 4,000	Coverage reached 87.99% Coverage reached 87.34%			
Dang, 2022b	EMI lectures Non-EMI lectures Open-access EMI courses	3,000 4,000 3,000	7,000 9,000 8,000	76.83% 80.47% 81.29%	6.40% 5.13% 5.12%	3.29% 2.64% 2.90%	0.51% 0.19% 0.51%

Coverage expressed in word families plus proper nouns and marginal words.

Table 1. A summary of lexical studies exploring general and academic coverage of specific academic genres.

From this Table, three aspects can be highlighted. First, most studies employ the BNC/COCA lists (Nation, 2012) as base lists to profile the corpus. These lists consist of 25 word-family lists arranged by frequency in addition to four supplementary lists comprising proper nouns (e.g., *Peter, Melissa*), marginal

words including swear words, exclamations, and letters of the alphabet (e.g., *aha, ehm*), acronyms and abbreviations (e.g., DPO, GDP), and compounds (e.g., *notebook, backpack*).

Second, some studies complement the study of lexical demands with the analysis of the coverage of the ASWL (Dang et al., 2017). This list comprises 1,741 word families clustered into four frequency levels. The first level includes 830 word families; level 2 consists of 456 word families; the third level is made up of 380 word families and level 4 has 75 word families. The word families in levels 1 to 3 are taken from the first, second, and third 1,000 frequency levels of the BNC/COCA, respectively. Level 4 includes words that are outside of high-frequency terms.

Third, studies can be classified based on whether English serves as the students' L1 or L2. Some studies have explored lexical demands in English L1 academic settings. For instance, Dang and Webb (2014) examined L1 lectures and seminars by determining the lexical load of the British Academic Spoken English (BASE) corpus, concluding that speakers needed to master the 4,000 and 8,000 most frequent word-families to reach 95% and 98% coverage, respectively. Similarly, Coxhead et al. (2017) studied the lexical demands of L1 small-group academic interactive settings by compiling two corpora: one on laboratory settings (137,399 running words) and another on tutorials (380,078 running words). They found that a vocabulary size of 3,000 word families plus proper nouns and marginal words provided 95% coverage in the laboratory corpus, and 98% coverage was reached with 7,000 word families. For the tutorial corpus, a vocabulary size of 2,000 words plus proper nouns and marginal words provided 95% coverage, and 98% coverage was reached with 3,000 word families plus proper nouns and marginal words.

A second set of studies has investigated general and academic vocabulary demands in English as a Lingua Franca in Academic settings (ELFA) or L2 linguistic contexts. Some studies have focused on the lexical demands of TED talks, such as Nurmukhamedov (2017) or Liu and Chen (2019), who analysed a corpus of 4.37 million words and the variations of lexical demands among topics. They found that recognising 3,000 and 6,000 word families provided 95% and 98% coverage, respectively, and that the ASWL covered nearly 90% of the corpus. Furthermore, they found that some topics may be less challenging for L2 learners than others. Other studies have concentrated their efforts on exploring other ELF academic fields. For instance, Dang

(2022a) analysed a corpus of 104 conference and colloquia presentations (565,758 tokens) to determine the general and academic lexical coverage. For general vocabulary, recognising the 3,000 most frequent word families provided 95% coverage, with the figure increasing to the 5,000 most frequent word families for 98% coverage. Regarding academic vocabulary, approximately 87.5% of the corpus was covered by the ASWL. In another study, Dang (2022b) compared the general vocabulary load of EMI, non-EMI L1 English, and open-access non-EMI courses, concluding that learners needed to recognise the 3,000 most frequent word families for 95% coverage in EMI and open-access non-EMI courses and up to the 4,000 band in non-EMI courses. For 98% coverage, the threshold was set at the 7,000 band for EMI courses, 8,000 band for the open-access non-EMI course, and 9,000 band for non-EMI courses.

Based on these studies, it can be assumed that the ASWL provides comprehensive coverage in a broad range of academic speaking scenarios and that understanding of the 3,000 most common English word families may enable comprehension of 95% of academic material. Nevertheless, the optimal level of comprehension may differ depending on the specific academic task. The minimal threshold also applies to EMI settings, where students seem able to access content with knowledge of a relatively low number of word families (Dang, 2022b). However, Dang's (2022b) study focuses exclusively on lectures, and before reaching sound conclusions on lexical demands on EMI, further research, contemplating other academic genres, such as seminars or tutorials, would be needed. As this author (2022a, p. 2) states, "the lexical demands of academic speech vary according to the type of speech events", and EMI courses, like any other academic event, usually include several activities varying in degree of interaction or linguistic demands.

The present study investigates vocabulary demands in EMI academic seminars, a genre that has gained recent recognition (Aguilar, 2016). Unlike lectures, which are mainly led by lecturers and can have a large audience, seminars have fewer participants and are prototypically characterised by being student-centred and dialogic in nature (Aguilar, 2016). The high degree of interaction in seminars can make active participation challenging, particularly for non-native English speakers (Aguilar, 2016; Jones, 1999). This may be especially true in EMI settings, where learners have varying levels of English proficiency and linguistic backgrounds. For this reason, this study aims to explore the lexical demands of EMI seminars.

3. Research questions

This study aims to enhance our understanding of EMI by determining the vocabulary demands in academic EMI seminars and exploring the impact of some contextual factors affecting this variable. This objective is defined in three research questions:

RQ1: What are the general vocabulary loads of EMI academic seminars? Do lexical demands in lecturers' and students' speech vary significantly?

RQ2: What is the coverage of the ASWL in EMI academic seminars?

RQ3: To what extent do contextual variables, such as classroom talk dominance or the lesson's topic, affect lexical coverage in EMI seminars?

4. Methodology

4.1. Corpus background information: MetCLIL

Data were obtained from the Corpus of Metaphor in Academic Talk (METCLIL), one of the primary outcomes of a Spanish national-funded project (Alejo et al., 2021) exploring the use of figurative language in EMI settings.

METCLIL is an open-access corpus containing the transcripts of nine academic EMI seminars on business and marketing and accounting for 111,061 tokens[1]. The transcripts were collected in six European Higher Institutions from Italy, the Netherlands, Norway, Portugal, Spain and Sweden. Seminars vary in duration, level of interaction, number of participants, and number of tokens uttered, as shown in Table 2. As can be seen from this table, some seminars (e.g., the Italian one) heavily relies on the lecturer's discourse, while others (e.g., the Swedish and Spanish seminars) are highly interactive, with students having a key role in the development of the lesson. Additionally, as displayed in Table 3, the topics covered in the seminars vary greatly, with some seminars focused on theoretical discussions about abstract concepts (e.g., the Dutch seminar), while others apply these concepts to real-world scenarios (e.g., the Norwegian or the Portuguese seminars). Table 3 further highlights the diverse range of topics covered in the seminars.

Setting	Length	No. turns / minute	No. of seminars	Overall no. of tokens	No. of tokens [lecturers]	No. of tokens [students]	Teacher talking time
Italy	108'	0.63	1	8,674	7,728	1,097	89%
Norway	120'	2.56	1	14,854	6,954	7,306	47%
The Netherlands	187'	2.68	1	21,117	11,800	9,794	55%
Spain	270'	4.00	3	34,324	13,124	22,141	38%
Portugal	80'	1.58	1	8,681	4,355	4,595	51%
Sweden	167'	10.97	2	22,843	5,816	17,032	25%

Table 2. Linguistic description of MetCLIL seminars.

4.2. Participants

METCLIL speakers are university lecturers and students attending Marketing and Business seminars taught in English in places where English does not hold official status. The whole corpus features 144 participants (six lecturers and 138 students) with varying L2 proficiency levels (ranging from B1 to C2) and more than 15 different L1s. However, not all the participants actively interacted in the seminars; thus, the information provided in this study will only refer to the total number of speakers (see Table 3).

Seminars	Topic	Role	N	No. of speakers	L1	L2 level
Italy	Marketing of typical products	Lecturer	1	1	Italian	C1
		Students	37	7	Romance: 4 Germanic: 1 Other: 2	B1: 1 B2: 3 C1: 3
Norway	Case study on group management	Lecturer	1	1	Norwegian	C1
		Students	17	12	Romance: 2 Germanic: 10	B2: 5 C1: 5 C2: 2
The Netherlands	Business management and its relationship with other fields	Lecturer	1	1	Dutch	C2
		Students	10	10	Romance: 1 Germanic: 6 Other: 2 Unspecified: 1	B2: 1 C1: 6 C2: 3
Spain	Marketing pitches	Lecturer	1	1	Spanish	C2
		Students	39	39	Romance: 25 Germanic: 9 Other: 4 Unspecified: 1	B1: 1 B2: 11 C1: 18 C2: 9
Portugal	Aid and development	Lecturer	1	1	Portuguese	C1
		Students	18	10	Romance: 7 Other: 3	B2: 2 C1: 7 C2: 1
Sweden	Multinational marketing strategies: interaction of marketing & culture	Lecturer	1	1	Chinese	B2
		Students	17	17	Romance: 3 Germanic: 1 Other: 13	B1: 1 B2: 4 C1: 7 C2: 4 Unspecified: 1

Table 3. Description of the seminars concerning their topic and speakers' features.

4.3. Data analysis

The corpus was pre-processed by removing some elements, such as truncated words, that were included in the original METCLIL version and could be disruptive to this study. The data was then analysed using AntWordProfiler v.2.0.1 (Anthony, 2022), a corpus analytic tool that allows the analysis of the lexis contained in a text based on specified word lists.

Two vocabulary lists were employed as base lists to profile the corpus: Nation's (2012) BNC/COCA lists and ASWL (Dang et al., 2017). In the case of the BNC/COCA lists, the additional lists were adapted to include proper nouns, abbreviations, acronyms, and foreign terms employed in METCLIL and not considered in the lists. As for the ASWL, this list was selected for two reasons: (1) it was created from a collection of academic spoken discourses, including seminars, which may help us determine if there are any differences between L1 and L2 seminars in terms of academic vocabulary, and (2) it was developed to aid L2 academic learners in developing their academic language skills by focusing on the most frequently used academic words.

5. Results

5.1. RQ1. What are the general vocabulary loads of EMI academic seminars? Do lexical demands in lecturers' and students' speech vary significantly?

The corpus analysis reveals that the first 1,000 word families account for nearly 87% of the items, followed by the second 1,000 (5.65%) and the third 1,000 word families (3.33%). Concerning off-list terms, i.e., terms not included in the BNC/COCA lists, they account for 0.41% of the whole corpus, most of which are L1 terms.

There is a relevant proportion of proper nouns, representing 1.33%, and marginal words, accounting for 0.73% of the whole corpus. Previous research (Dang, 2022a; Dang & Webb, 2014; Webb & Rodgers, 2009a, 2009b) indicated that these kinds of words have a lower learning burden than other content words, and the current research adhered to their guidance. In practice, this means that 95% coverage is reached with knowledge of the 2K most frequent terms plus additional terms, whereas 98% coverage is reached with the 3,000 most frequent family words and additional terms.

Table 4 provides the cumulative coverage of the fifth 1,000 word families and the upper limit of mid-frequency (8,000 word family band) terms. It also details the figures of proper nouns, abbreviations and acronyms (A&A), marginal words and off-list words.

Word families	Overall		Lecturers		Students	
	Cumulative coverage (CC)	Cumulative coverage plus additional lists (CC+)	Cumulative coverage (CC)	Cumulative coverage plus additional lists (CC+)	Cumulative coverage (CC)	Cumulative coverage plus additional lists (CC+)
1,000	86.49	89.19	87.89	89.61	86.35	88.65
2,000	92.53	95.23[a]	93.14	94.95	92.47	94.77
3,000	95.81	98.52[b]	95.91	98.19[a,b]	95.81	98.11[a,b]
4,000	96.37	99.07	96.41		96.36	98.67
5,000	96.73	99.43	96.67		96.74	99.04
8,000	97.09	99.82	96.95		97.13	99.44
Off-list	0.41		0.62		0.38	
Proper nouns	1.33		1.10		1.35	
Marginal words	0.72		0.86		0.71	
A&A	0.12		0.17		0.12	
Compounds	0.12		0.04		0.13	

a 95% coverage reached.
b 98% coverage reached.

Table 4. Coverage of Nation's (2012) BNC/COCA lists in the METCLIL corpus.

Finally, the lexical production of lecturers and students was compared and showed little difference, with both groups achieving 95% and 98% coverage using the 3,000 most frequent word families.

5.2. What is the coverage of the aswl in EMI academic seminars?

Overall, the ASWL annotated 76.22% of the METCLIL corpus, with 70.85% falling into level 1, 3.54% into level 2, 1.72% into level 3, and only 0.11% into level 4.

Concerning the comparison of lecturers' and students' ASWL coverage, lecturers demonstrate a slightly higher use of academic terms situated in levels 1, 3 and 4, whereas students tend to have a greater proportion of level 2 word families in their speech.

Level	No. of word families	Overall (in %)	Lecturers (in %)	Students (in %)
Level 1	830	70.85	71.12	70.63
Level 2	456	3.54	3.28	3.74
Level 3	380	1.72	1.75	1.70
Level 4	75	0.11	0.13	0.09

Table 5. Coverage of Dang et al. (2017) ASWL lists in the METCLIL corpus.

5.3. RQ3: To what extent do context variables, such as classroom talk dominance or lesson topic, affect lexical coverage in EMI seminars?

Research question 3 aimed to explore whether two contextual variables (the topic of the seminars and classroom talk dominance) may affect lexical coverage.

Concerning topics, METCLIL contained data from nine seminars in six different learning situations and focused on various marketing and business topics. Thus, the vocabulary load was calculated per learning situation to assess how the topics impacted the seminars' lexical demands.

As seen from Table 6, there appear to be differences in the vocabulary demands among the different learning contexts. The 2,000 word families, along with the additional lists, achieve 95% coverage in the Spanish and Swedish seminars, whereas, for the rest of the seminars, comprehension of the 3,000 most frequent word families plus the additional lists is needed. To reach very high comprehension (98% coverage), knowledge of the 3,000 most frequent word families and additional terms suffices for the Norwegian, Portuguese, and Swedish seminars. On the other hand, the Italian and Spanish seminars require an understanding of the 4,000 most frequent word families and the additional lists for optimal comprehension, while the Dutch seminar demands a mastery of the 5,000 most frequent word families along with the additional lists.

The exploration of the words belonging to additional lists in isolation also identifies some coverage pattern differences. On the one hand, the employment of compounds and abbreviations and acronyms are, for their part, anecdotal. On the other hand, proper nouns and marginal words are more prominent in the corpus. Proper nouns, serving as vocatives, contextual markers (referring to places, companies or people), and theoretical references, constitute 1.33% of the corpus. However, their usage varies widely among seminars, ranging from 0.60% in the Italian seminar to 2.04% in the Spanish seminar. Likewise, marginal words have a consistent, albeit lower, frequency in the seminars with varying degrees of coverage: they make up 0.20% of the total vocabulary in the Norwegian seminars and 2.44% in the Dutch seminar.

Learning context		1,000	2,000	3,000	4,000	5,000	8,000	Off-list	Proper nouns	Marginal words	Compounds	A&A
Italy	CC	85.01	93.45	96.46	96.73	97.04	98.06	0.31	0.60	0.55	0.09	0.19
	CC+	86.45	94.89	**97.90**[a]	**98.16**[b]	98.47	99.50					
The Netherlands	CC	82.65	88.59	94.11	94.94	96.12	96.52	0.18	0.17	2.44	0.09	0.26
	CC+	85.61	91.55	**97.07**[a]	97.90	**99.08**[b]	99.48					
Norway	CC	87.55	92.66	96.70	97.57	97.74	97.84	0.51	0.95	0.20	0.29	0.09
	CC+	89.08	94.19	98.23b	99.11	99.27	99.38					
Portugal	CC	85.94	93.32	97.11	97.86	98.28	98.44	0.13	1.07	0.17	0.10	0.06
	CC+	87.34	94.72	**98.55**[a,b]	99.26	99.68	99.84					
Spain	CC	86.47	92.33	95.06	95.56	95.83	96.26	0.62	2.04	0.65	0.09	0.14
	CC+	89.39	**95.25**[a]	97.98	**98.49**[b]	98.75	99.18					
Sweden	CC	89.22	94.58	97.03	97.35	97.54	97.84	0.13	1.03	0.60	0.11	0.03
	CC+	91.00	**96.36**[a]	**98.80**[b]	99.12	99.32	99.62					

a 95% coverage reached.
b 98% coverage reached.

Table 6. Coverage of Nation's (2012) BNC/COCA lists by METCLIL seminars.

Similarly, the coverage of Dang et al.'s (2017) ASWL was examined in each seminar context to explore variations in academic vocabulary. As seen in Table 7, no large and systematic differences are observed. For Level 1, the Norwegian, Spanish, and Swedish seminars showed higher coverage than the Italian, Dutch, and Portuguese seminars. For level 2, coverage in the Italian and Portuguese seminars showed larger coverage when compared to the Dutch, Swedish, Norwegian and Spanish seminars, in which the coverage was substantially lower. Level 3 coverage showed substantial differences, with the Italian, Dutch, and Norwegian seminars having twice the coverage found in the Portuguese, Spanish, and Swedish seminars.

Level	No. of word families	Overall (in %)	Italy	The Netherlands	Norway	Portugal	Spain	Sweden
Level 1	830	70.85	67.64	68.06	74.17	70.69	71.97	70.96
Level 2	456	3.54	5.10	3.21	3.3	5.27	2.97	3.6
Level 3	380	1.72	2.73	2.8	2.14	1.2	1.09	1.23
Level 4	75	0.11	0.09	0.35	0	0.04	0.07	0.05

Table 7. Coverage of Dang et al.'s (2017) ASWL by seminars.

At first, the results seem to indicate that the content of the seminars somewhat influences the lexical demands of the seminars, and this is evident in some particular situations. For instance, the Dutch seminar exhibits the most significant disparities compared to other situations, with lower coverage of high-frequency terms from the BNC/COCA lists and the ASWL, a higher proportion of less common words, and reduced reliance on proper nouns. This seminar focuses on more abstract and theoretical concepts

(analogical reasoning, borrowing and blending as marketing tools) than other learning contexts that include concrete examples, personal experiences, or case studies, potentially contributing to these differences.

The seminars not only shifted in terms of classroom topic but also based on the dominance of the classroom discourse. As shown in Table 2, the seminars can be classified into three categories based on their teacher talking time: lecturer-led (Italian seminar), balanced (Norwegian, Portuguese, and Dutch seminars), and student-led (Spanish and Swedish seminars). This classification was used to assess the impact of this variable on the vocabulary demands of seminars.

Data indicate that seminars where students' speech plays a major part result in a lower vocabulary load, with comprehension reaching 95% and 98% for the 2,000 and 3,000 most frequent word families and additional lists, respectively. Conversely, in seminars where lecturers have a more prominent role, or where both students and lecturers participate equally, relatively high and stable comprehension (95% coverage) is achieved through knowledge of the 3,000 most frequent word families and additional lists, while high-level comprehension (98% coverage) is attained with knowledge of the 4,000 most frequent word families and additional lists.

Besides, some differences are also spotted when focusing on additional lists concerning proper nouns and off-list coverage. As can be seen from Table 8, student-led interaction seminars present a substantially higher proportion of off-list terms and proper nouns, while balanced interaction seminars display greater employment of marginal words.

Word families	Lecturer-led interaction		Balanced interaction		Students-led interaction	
	CC	CC+	CC	CC+	CC	CC+
1,000	85.01	86.45	85.49	87.44	87.26	89.85
2,000	93.45	94.89	91.54	93.49	92.98	**95.56**[a]
3,000	96.46	**97.90**[a]	95.99	**97.95**[a]	95.62	**98.21**[b]
4,000	96.73	**98.16**[b]	96.80	**98.76**[b]	96.07	98.66
5,000	97.04	98.47	97.37	99.33	96.32	98.91
8,000	98.06	99.50	97.59	99.54	96.72	99.30
Off-list	0.31		0.29		0.48	
Proper nouns	0.60		0.73		1.75	
Marginal words	0.55		0.92		0.63	
A&A	0.19		0.14		0.11	
Compounds	0.09		0.10		0.10	

a 95% coverage reached.
b 98% coverage reached.

Table 8. Coverage of Nation's (2012) BNC/COCA lists according to classroom talk dominance.

Finally, ASWL coverage was also examined to check whether classroom talk dominance affected academic vocabulary significantly. As seen from Table 9, a marked pattern of ASWL usage seems to be in correlation with the proportion of teacher talking time. As teacher talking time decreases, the coverage of level 1 word families rises and the coverage of level 2 and level 3 word families diminishes.

Level	No. of word families	Lecturer-led interaction	Balanced interaction	Students-led interaction
Level 1	830	67.64	70.53	71.58
Level 2	456	5.10	3.65	3.21
Level 3	380	2.73	2.27	1.15
Level 4	75	0.09	0.18	0.06

Table 9. Coverage of Dang et al. (2017) ASWL according to classroom talk dominance.

6. Discussion

The literature suggests that lexical coverage acts as one of the main predictors of L2 comprehension (Laufer & Sims, 1985). Research has demonstrated that 95% and 98% coverage result in very high comprehension and very high and stable comprehension, respectively (Vilkaite-Lozdiene & Schmitt, 2020). Furthermore, research has provided information on the number of word families required to reach this coverage in academic settings, indicating that recognition of the top 3,000 English word families results in 95% coverage, while 98% coverage can be attained through comprehension of 4,000 to 8,000 word families.

However, most studies have focused on educational settings where the language of instruction is the mother tongue of the majority of learners (Dang & Webb, 2014; Coxhead et al., 2017), a path that has yielded valuable insights into the lexical demands in anglophone academic contexts for L2 learners. Nevertheless, there is a growing trend in non-anglophone countries to offer a part of their academic curriculum in a second language, predominantly English, and this calls for further investigation into the linguistic demands of these specific learning contexts.

This study examined the vocabulary demands of academic EMI seminars, a common type of educational activity in EMI settings. While previous research has found that EMI lectures may place lower lexical demands than their non-EMI counterparts (Dang, 2022a), the vocabulary load can still vary significantly depending on the specific educational event. To better

understand the linguistic demands of EMI settings, this study focused on academic seminars, an educational genre that often involves dialogic interaction and a more active role for learners. This can pose difficulties for L2 learners, as effective participation may affect language usage. By exploring the vocabulary load of EMI seminars and the impact of contextual variables, this study has explored the linguistic demands faced by L2 learners in EMI seminars, explicitly focusing on lexical coverage of General and Academic English lists (BNC/COCA and ASWL, respectively), and the impact of contextual variables (i.e., classroom talk dominance and lesson' topic) on lexical coverage in EMI.

6.1. BNC/COCA coverage

Research Question 1 aimed to assess the general vocabulary demands of EMI academic seminars, using Nation's (2012) BNC/COCA list to this aim. The results revealed that the 2,000 most frequent word families and the additional lists covered 95% of the corpus, while 98% coverage was achieved with knowledge of the 3,000 most frequent word families plus additional lists. Furthermore, marginal use of low-frequency terms (+9,000 word families) was identified. These findings align with previous research in academic tutorials, which also found that 95% and 98% coverage could be obtained through recognition of the 2,000 and 3,000 most frequent word families, respectively (Coxhead et al., 2017), and with research in academic spoken English, which suggests limited use of low-frequency word families in academic spoken contexts (Coxhead et al., 2017; Dang, 2022a, 2022b; Dang & Webb, 2014). Conversely, the lexical demands of EMI seminars are lower than those of other academic settings, such as L1 lectures, seminars, and TED talks, where 95% and 98% coverage is achieved with knowledge of the 4,000 and 8,000-9,000 word families plus additional terms, respectively. This difference can be attributed to two factors. Firstly, lectures often convey a large amount of technical or academic information, requiring mid- and low-frequency items. Secondly, lectures and TED talks are more monologic, while EMI seminars have a diverse audience, comprising mostly L2 students who actively participate in the class. This diversity may lead to simpler and more accessible language to promote effective communication among peers.

Furthermore, this study has specifically examined the use of proper nouns, off-list terms and marginal words in isolation. These terms are often overlooked in lexical coverage studies as they are believed to have a low learning burden (Webb & Rodgers, 2009a, 2009b). However, the analysis of

the coverage distribution by seminars revealed that these terms varied greatly depending on the seminars and, consequently, further research was needed. Data suggests a slightly heavier employment of proper nouns and off-list words than in L1 tutorials (Dang et al., 2017) and a lower usage of off-list terms compared to L1 laboratories. Concerning the variations in off-list words, their increased use may be attributed to the nature of the METCLIL corpus, which primarily consists of L2 speakers from diverse linguistic backgrounds using English as a lingua franca. Unlike L1 corpora, where off-list words are usually low-frequency words, in the present study, off-list words are mostly speakers' L1 words. Some students seem to resort to their mother tongue when they struggle to communicate effectively in English or feel misunderstood. This phenomenon is especially evident in peer-interaction activities, and it may be a particularity of EMI speech that highlights the importance of reconsidering the exploration of off-list words in lexical coverage studies, especially if the corpus includes English L2 speakers.

As for proper nouns, a closer examination of their roles in METCLIL displays three predominant functions: as vocatives (primarily found in highly interactive seminars), to contextualise information (referring to places, companies or people, equally found in most seminars), and to refer to theoretical points (mainly authors of business theories or papers). As some of the participants in METCLIL were international students, specific proper nouns referring to local places near the academic institution may pose difficulties for them due to their context dependence. If this occurs, EMI lecturers should be aware of this situation and try to scaffold the understanding of these proper names as well.

The second part of RQ1 explored the variation in lexical demands of lecturers' and students' speech. The results indicated minimal variations, except for abbreviations and acronyms, which are significantly more used in lecturers' speeches. These abbreviations and acronyms are often specific to particular fields of study and may still be unfamiliar to students who are in the process of acquiring their meanings in the L2. The similarities between the lexical demands of students' and lecturers' speeches can result from two different situations. First, lecturers may adapt their speech style to promote comprehension, leading to lower lexical demand (Mauranen, 2012). Second, in certain seminars, students assumed the role of lecturers and provided feedback to their peers, which may have affected their language use. Further research would be welcome to investigate the extent to which students imitate and incorporate lecturers' feedback into their own language use.

6.2. ASWL coverage in EMI seminars

RQ2 examined the occurrence of academic vocabulary in METCLIL employing the ASWL (Dang et al., 2017) as a base list. The ASWL was deemed appropriate for this purpose due to its demonstrated representation of the distinctive features of academic spoken vocabulary. Derived from a corpus of 13.5 million words that included four different academic spoken genres, the ASWL was intended to serve as a benchmark for evaluating its prevalence in METCLIL compared to other academic contexts.

In general, the ASWL accounted for 76.22% of the METCLIL corpus. Level 1 comprised 70.85% of the whole corpus, while levels 2, 3 and 4 amounted to 3.54%, 1.72% and 0.11%, respectively. This coverage was lower than reported in other studies (Dang et al., 2017; Dang, 2022a, 2022b; Liu & Chen, 2019), particularly for levels 1 and 2. This difference could be partly attributed to two factors. First, the speakers in METCLIL are second language learners, and previous research (Dang, 2022b) has already indicated that EMI academic settings tend to offer lower ASWL coverage than comparable educational settings where English native students receive instruction in their L1. This study confirms prior findings. Secondly, the present study only focused on EMI seminars, a genre linguistically characterised by dialogic interactions in which students take an important role. This situation may lead to the use of more everyday language and a reduction of academic and specialised terminology, although further research is needed before drawing firm conclusions.

In light of these results, EMI students are less prone to acquire academic spoken vocabulary from these seminars in comparison to other academic environments such as TED talks, English L1 or EMI lectures. This observation supports the previously stated idea by Dang (2022a) that the relevance of the ASWL for second language learners may vary depending on the context. However, as there is limited research on the ASWL in L1 seminars, further studies are needed to determine to what extent the ASWL coverage in EMI seminars differs from similar educational settings in terms of interaction and student participation.

6.3. Contextual variables and lexical demands

Research Question 3 aimed to investigate how two contextual variables (lesson topic and involvement in classroom talk) affect EMI seminars' lexical demands. Prior studies exploring aboutness in TED talks (Liu & Chen, 2019)

found that the topic influenced coverage levels to some extent. The current piece of research partially supports this finding, as the results show that some topics seem more lexically demanding than others. This is particularly evident in the case of the Dutch context, where a larger proportion of mid-frequency words are needed to reach minimal and optimal coverage, or in the Swedish context, where lower lexical demands are found. However, our analysis revealed that some of these differences could also be linked to the level of interaction and teacher talking time in the seminars. Interestingly, the results suggest a negative relationship between student involvement and seminars' lexical demands. For instance, the Swedish and Spanish seminars, which have the highest level of student participation, showed the lowest lexical demands for very high comprehension (95% coverage). The Communicative Accommodation Theory (see Gallois et al., 2005, for a general overview, and Mao, 2006; Park & Kang, 2010, for its application in L2 classrooms) offers a plausible explanation for this phenomenon. Its main tenets posit that speakers adopt accommodative strategies and converge with others to enhance mutual understanding. EMI learners have a wide variety of L2 proficiency levels ranging from B1 to C2, but still, need to communicate to accomplish tasks. Thus, students with higher L2 proficiency may be resorting to accommodative measures to bridge the language proficiency gap with their less proficient peers, although further research is needed to draw conclusive findings.

However, the seminar in the Netherlands stands as an exception. Despite having a balanced interaction between the lecturer and the students, it is the most lexically demanding seminar, reaching 98% at the highest word family levels. In this specific case, it could be argued that the content of the seminar may play a significant role, as the topic discussed in the Dutch seminar revolves around more abstract and theoretical ideas compared to the rest of the seminars that concentrate on concrete examples, personal experiences, or case studies. Abstract topics often require the employment of more specialised terminology and abstract terms, which are less frequently found in oral and written communication. This phenomenon could be impacting the lexical demands of this specific learning context.

In other words, while lesson topics may play a role in lexical demands, classroom talk dominance seems to impact general vocabulary demands, which also seems to hold for academic vocabulary. Specifically, our findings show that as the percentage of teacher talking time increases, seminars tend to use more Level 2 and 3 academic words while using fewer Level 1 terms.

These results suggest that the degree of teacher involvement in the learning process may be critical for shaping students' acquisition and use of academic vocabulary. However, further research is needed to fully understand the relationship between teacher talk dominance, lesson topic, and general and academic vocabulary coverage in EMI seminars.

7. Conclusions

This study has explored the general and academic vocabulary demands in a corpus of EMI seminars and the impact of contextual factors (topic and teacher talking time) on this variable. With the increasing popularity of EMI programmes worldwide, there is a growing interest in understanding the linguistic features of this type of learning situation and the linguistic and lexical demands they place on L2 learners, and this piece of research aims to address this gap.

The results have shown that, to achieve 95% and 98% coverage, EMI seminars require knowledge of the 2,000 and 3,000 most frequent lexical items, respectively, plus additional lists. Compared to other academic settings, EMI seminars entail a lower lexical complexity for students, which implies that learners may not perceive this learning scenario as linguistically demanding as English L1 seminars, tutorials, or TED talks based on vocabulary alone. This finding is supported by the analysis of academic lexical coverage, which demonstrates that EMI seminars contain a smaller proportion of academic terms than EMI lectures (Dang, 2022b), L1 lectures (Dang, 2022b; Dang et al., 2017), or TED talks (Liu & Chen, 2019).

The findings also suggest that certain contextual factors, such as the degree of interaction and the percentage of lecturer talking time, can influence the lexical demands of EMI seminars. Specifically, when students are given a more active role in this learning situation, there is a noticeable reduction in lexical complexity. On the other hand, the impact of other contextual variables, such as the seminar content, is less clear, and further research is needed to determine the extent to which the seminar topic would account, in and by itself, for variations in lexical demands.

While this study provides valuable insights, it is important to acknowledge its limitations. The first one is the scope of the corpus analysed, which is limited in size and only comprises seminars discussing issues related to Marketing and Business. This research could be more representative if a more extensive

set of seminars from various fields were analysed. Second, the general lexical demands of spoken language were identified using the BNC/COCA lists as a base list. However, these lists were primarily derived from written and spoken texts, which may not accurately capture the unique features of spoken language. Therefore, future research should employ frequency lists exclusively sourced from spoken corpora to provide a more comprehensive understanding of the lexical demands of spoken language. Thirdly, while the METCLIL corpus included speakers with diverse linguistic backgrounds and L2 proficiency levels, the impact of L2 proficiency on lexical coverage was not considered. To the best of our knowledge, no previous research has examined this aspect, and investigating this aspect could provide valuable insights for EMI research. In this respect, it would be relevant to include information about learners' receptive and productive vocabulary knowledge through the use of validated tests. Doing so would enhance our understanding of learners' L2 proficiency. Finally, the study exclusively focused on the analysis of spoken transcripts. Including other methods, such as interviews or questionnaires, to triangulate the results would enrich the findings.

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics, 24*, 425-438.

Aguilar, M. (2016). Seminars. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for Academic Purposes* (pp. 337-349). Routledge.

Alejo, R., Piquer-Píriz, A. M., Castellano-Risco, I., Martín-Gilete, M., Fielden-Burns, L., MacArthur, F., Nacey, S., Philips, G, Krennmayr, T., Coelho, M., Littlemore, J., & Ädel, A. (2021). METCLIL *corpus v1*. https://www.sketchengine.eu/metclil-corpus-of-metaphor-in-academic-talk/

Anthony, L. (2022). *AntWordProfiler* (Version 2.0.1) [Computer Software]. Waseda University. https://www.laurenceanthony.net/software

Bauer, L., & P. Nation. (1993). Word families. *International Journal of Lexicography, 6*(4), 253-279. https://doi.org/10.1093/ijl/6.4.253

Beelen, J., & Jones, H. (2015). Redefining internationalization at home. In R. Pricopie, J. Salmi, P. Scott & A. Curai (Eds.), *Redefining internationalization at home* (pp. 67-80). Springer.

Coelho, D. (2022). Plurilingual pedagogy in higher education in the UAE: Student voices in an academic writing course. In D. Coelho & T. G. Steinhagen (Eds.), *Plurilingual pedagogy in the Arabian Peninsula* (pp. 95-113). Routledge. https://doi.org/10.4324/9781003315971

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, *34*(2), 213-238.

Coxhead, A., Dang, T.N.Y., & Mukai, S. (2017). Single and multi-word unit vocabulary in university tutorials and laboratories: Evidence from corpora and textbooks. *Journal of English for Academic Purposes, 30*, 66-78. https://doi.org/10.1016/j.jeap.2017.11.001

Dang, T. N. Y. (2022a). A corpus-based study of

vocabulary in conference presentations. *Journal of English for Academic Purposes, 59*, 101144. https://doi.org/10.1016/j.jeap.2022.101144

Dang, T. N. Y. (2022b). Vocabulary in academic lectures. *Journal of English for Academic Purposes, 58*, 101123. https://doi.org/10.1016/j.jeap.2022.101123

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *Journal of English for Academic Purposes, 33*, 66-76. https://doi.org/10.1016/j.esp.2013.08.001

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning, 67*(4), 959-997.

Durrant, P., Siyanova-Chanturia, A., Kremmel, B., & Sonbul, S. (2022). *Research methods in vocabulary studies.* John Benjamins.

Gallois, C., Ogay, T., & Giles, H. (2005). Communication accommodation theory: An interdisciplinary framework for the study of adaptation in communication. In H. Giles, J. Coupland & N. Coupland (Eds.), *Handbook of language and social psychology* (pp. 3-26). John Wiley & Sons Ltd.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35,* 305-327. https://doi.org/10.1093/applin/amt015

Hellekjaer, G.O. (2010). Lecture comprehension in English-medium higher education. *HERMES-Journal of Language and Communication in Business, 45,* 11-34. https://doi.org/10.7146/hjlcb.v23i45.97343

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8*, 689-696.

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*, 403-430. https://doi.org/10125/66973

Jones, J. (1999). From silence to talk: Cross-cultural ideas on students participation in academic group discussion. *Journal of English for Specific Purposes, 18*(3), 243-259. https://doi.org/10.1016/S0889-4906(97)00059-8

Laufer, B., & Ravenhorst-Kalovski, G.C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*, 15-30. https://doi.org/10125/66648

Laufer, B., & Sim, D. D. (1985). Measuring and explaining the reading threshold needed for English for academic purposes texts. *Foreign Language Annals, 18*(5), 405-411.

Liu, C.Y., & Chen, H.H.J. (2019). Academic spoken vocabulary in TED talks: Implications for academic listening. *English Teaching & Learning, 43*, 353-368. https://doi.org/10.1007/s42321-019-00033-2

Macaro, E. (2018). *English medium instruction: Content and language in policy and practice.* Oxford University Press.

Mao, L. (2006). Communication accommodation in English as a lingua franca interactions. *Journal of Pragmatics, 38*(3), 437-460.

Mauranen, A. (2012). *Exploring ELF: Academic English shaped by non-native speakers.* Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*, 59-82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2012). *The BNC/COCA word family lists.* https://www.victoria.ac.nz/__data/assets/pdf_file/0004/1689349/Information-on-the-BNC_COCA-word-family-lists-20180705.pdf

Nation, I. S. P. (2013). *Learning vocabulary in another language* [2nd ed.]. Cambridge University Press.

Nurmukhamedov, U. (2017). Lexical coverage of TED talks: implications for vocabulary instruction. *TESOL Journal, 8*(4), 768-790. https://doi.org/10.1002/tesj.323

Park, H. J., & Kang, O. (2010). Communication accommodation in L2 interaction: The case of Korean EFL learners. *System, 38*(4), 607-620.

Pérez-Cañado, M.L. (2021). CLIL-ising EMI: An analysis of student and teacher training needs in monolingual contexts. In C. Hemmi & D.L. Banegas (Eds.), *International Perspectives on CLIL* (pp. 171-191). Springer.

Piquer-Píriz, A.M. (2023). *E-learning in English Medium Instruction (EMI): Academic language for university students.* Peter Lang.

Piquer-Píriz, A. M., & Castellano-Risco, I. O. (2021). Lecturers' training needs in EMI programmes: Beyond language competence. *Alicante Journal of English Studies, 34*, 83-105. https://doi.org/10.14198/raei.2021.34.05

Schmitt, N. (2010). *Researching vocabulary: A Vocabulary Research Manual.* Palgrave Press.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484-503. https://doi.org/10.1017/S0261444812000018

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, *95*(1), 26-43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Shepard, C., & Morrison, B. (2021). Challenges of English-medium higher education. The first-year experience in Hong Kong revisited a decade later. In D. Lasagabaster & A. Doiz (Eds.), *Language use in English-medium instruction at university* (pp.167-192). Routledge.

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*, 457-479. https://doi.org/10.1093/applin/ams074

Vilkaite-Lozdiene, L., & Schmitt, N. (2020). Frequency as a guide for vocabulary usefulness: High-, mid-, and low-frequency words. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 81-96). Routledge.

Webb, S., & Rodgers, M. (2009a). Vocabulary demands of television programs. *Language Learning, 59*(2), 335-366. https://doi.org/10.1111/j.1467-9922.2009.00509.x

Webb, S., & Rodgers, M. (2009b). The lexical coverage of movies. *Applied Linguistics*, *30*(3), 407-427. https://doi.org/10.1093/applin/amp010

Xue, G., & Nation, N. (1984). A university word list. *Language Learning and Communication, 3*(2), 215-229.

Irene Castellano-Risco is a Lecturer at the Faculty of Education and Psychology of the University of Extremadura, Spain, where she teaches undergraduate courses on EFL and Second Language Acquisition. Her research interests concern metaphorical and L2 lexical competence development, and the impact of content-based instruction (CLIL/ EMI programmes) on L2 learning. She has participated in several research projects financed by the Spanish Ministry of Education and the regional authorities in Extremadura. Her main publications have been published in journals such as *System*, *Onomázein, ES Review*, or *Journal of English Studies.*

NOTES

[1] Official METCLIL version released in Sketch Engine accounts for 110,496 tokens. Discrepancies in the total amount of tokens are related to how multi-word units are counted and to the deletion of truncated words in this study. In METCLIL public version, multi-word units (for example, "because of", "so as" or "such as") have been hyphenated and considered single lexical items. On the contrary, in this study, each element of multi-unit words is counted independently.