

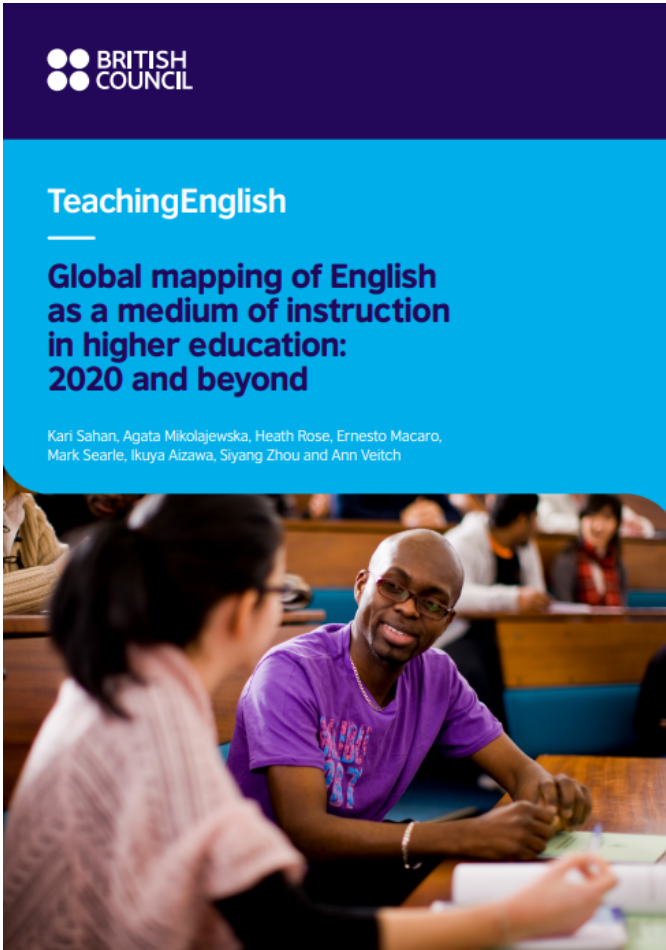
Developing the EMI Corpus of Student Academic Reading & Writing: Addressing challenges in a large-scale, transnational collaborative project

Dana Gablasova - Raffaella Bottini - Vaclav Brezina - Luke Harding - Sally Ren
Lancaster University

Overview of this session

1. EMI Corpus project: introduction
2. Overview of the EMI corpus
3. Challenges in the EMI Corpus multi-site collaborative project

English medium instruction (EMI)



- EMI – teaching/learning disciplinary subjects through the medium of English typically in countries where English is not the community language (McKinley, 2024; Pecorari & Malmström, 2018).
- EMI – currently a global pedagogical trend; on the increase
- Use and knowledge of English – crucial for understanding subject knowledge and for learning

EMI: Challenges related to language use

- We know a lot about EMI – reported via surveys, interviews, classroom observations, document analysis
- We know that students report **difficulties** related to speaking, writing and reading English – with potentially negative consequences for their academic success
- However, we do not have much data about how they **actually use English** and what **demands** are placed on them (e.g. in their reading) → **calls for corpus research in EMI** (Jablonkai, 2021)

EMI Corpus project



Project: “Linguistic demands of EMI in Higher Education: A corpus-based analysis of student writing and reading in EMI university settings.”

Funded by the British Council as part of the **Future of English research scheme** for 2022-25



Prince of Songkla University



Thammasat University



University of Turin



University of Milan



Xi'an Jiaotong University



Xi'an Jiaotong-Liverpool University



Vienna University of Economics and Business

Answers to the question no :- 3

Corpora of EMI language use



Description of linguistic patterns and regularities



Understanding what language students produce and encounter



Understanding of student challenges and needs



Inform language teaching and testing practice/materials; Inform EMI policy (e.g., admission requirements, EAP provision, ESP provision)

Corpus evidence and EMI

MA International Relations - MSc Computer Science – MBA - MSc Project Management - MA Intercultural Communication - MEng Mechanical Engineering - MA History – MBA - MA History - MSc Business Analytics - MSc Advanced Marketing Management - MEng Hons Chemical Engineering - MSc Finance - MSc Developmental Psychology - MSc Digital Business, Innovation and Management - MSc Human Resources Management - MA International and Military History - MSc Computing - MSc Mechanical Engineering - MA Creative Writing with English Literature - MA in History - PGCert Regional and Local History - MA Philosophy and Religion - BEng Chemical Engineering - MSc in Engineering - MA Language and Linguistics- BEng

Corpus of EMI reading and writing: Overview

MA in Leadership and Management - MA Management Science - MSc Developmental Disorders - MA Language and Linguistics - MSc Digital Business Innovation and Management - BEng Hons Mechatronic Engineering - MSc Advanced Mechanical Engineering - MA Applied Linguistics - MA Applied Linguistics and TESOL - MSci Biomedicine - MSc Entrepreneurship and Innovation - MA Politics - MA Digital Humanities - MSc Criminology and Social Research Methods - MA Digital Humanities - MEng Mechanical Engineering - MA Linguistics - MSc Biomedicine - MSc Advanced Marketing Management - MChem Chemistry Hons - MEng Mechanical Engineering - MA Media and Cultural Studies - MSc Business Analytics - MA English Literary Studies - MA Creative Writing (Distance Learning) - MA Criminology and Criminal Justice - MA Discourse Studies - MSc Advanced Mechanical Engineering - MA Social Justice and Education - MA Philosophy - MSc Health Research - MA Intercultural Communication - MSc Conservation and Biodiversity - MSc Volcanology and the Environment

Corpus size: 4.5M words and 2,000+ student papers

Level	Undergraduate		Postgraduate	
Disciplinary area	Business & Management		Humanities & Soc. Science	Science & Technology
Core subjects	Business studies, Management, Finance, Accounting,		History, Literature, Sociology, Linguistics	Chemistry, Physics, Engineering
Corpus size (Italy, Thailand, China, Austria)	234	283	525	345
Corpus size (UK-based data)	20	140	257	200
Corpus size (data total)	254	423	782	545

Frequency information

What downtoners are used in different disciplinary areas?

#LancsBox X 4.0.0

Q DOWNTONER

EMI_HUM

whole corpus

255K

Q DOWNTONER

Hits: 461 (1,809.77)

Texts: 46/50

File	Left	Node	Right
R_8PhQzb3cncrSpFl	's creation was a common approach, but not the	only	one. Averroism, a philosophical movement that caused controversy by
R_8PhQzb3cncrSpFl	such as Greek, Jewish, and Islamic works. It's	somewhat	accurate to say he aimed to somehow <footnotes>3
R_31bzYpQgGEjTv:	the same. Speaker1 So like the price, it's	almost	same in UK? like, for example, the apple, the
R_31bzYpQgGEjTv:	bonds when criticising UK food (Lines 1 to 9),	partly	because both are "novices" and foreigners in the local
R_8PhQzb3cncrSpFl	a perfect being, he reasons out that it can	only	come from a perfect source—that is, God. Accordingly,
R_8PhQzb3cncrSpFl	of such a being, then it must exist not	only	in the mind but also in reality. His reasoning
R_2D6Mldyn1Wg6	perhaps there are one or two definitions that are	barely	used. Next, ask ChatGPT to list out its results
R_2D6Mldyn1Wg6	context information to solve specific tasks, which can not	only	improve the effects of various tasks, but also better

EMI_ENG

whole corpus

358K

Q DOWNTONER

Hits: 396 (1,105.46)

Texts: 51/61

File	Left	Node	Right
R_2iEfknJ	Catalysed Deuterium (SCD) reaction, where	only	tritium is burnt in a
R_2iEfknJ	the overall resultant reaction, producing	only	helium. <formula/> Breeding
R_2iEfknJ	carried by charged particles. Not	only	does this have a benefit
R_2iEfknJ	stated that people can generally	only	hold 7 ± 2 objects
R_2iEfknJ	important to not rely on	only	an FMA to generate the
R_2iEfknJ	a restricted symbology consisting of	only	two shapes and one type
R_2iEfknJ	could lead to related, but	slightly	more realistic, ideas. A graphical
R_2iEfknJ	as simple as possible so	only	the broad functions are given,
R_2iEfknJ	despite being more complex and	slightly	more expensive. It utilizes multiple

EMI_BUS

whole corpus

156K

Q DOWNTONER

Hits: 174 (1,112.50)

Texts: 40/50

File	Left	Node	Right
R_2Sk7SS	a business change, they can	hardly	survive in their initial operation
R_2Sk7SS	2017); Rolex (n.d.) By knowing	only	four elements of Rolex's
R_2Sk7SS	slowly shifted their game to	almost	fully in-house production with
R_2Sk7SS	change at once One-way,	only	one operator can be implemented
R_2Sk7SS	that innovating a business model	only	becomes applicable when a business
R_2Sk7SS	cliche. At the time, the	only	examples I was exposed to
R_2n3UsC	and social considerations is not	merely	a trend but a strategic
R_8du2n	the 'economy of simplicity' not	only	helps companies simplify their operations
R_8du2n	assembly. This smart strategy not	only	saves time since it does

What adjectives are common in different disciplines?

#LancsBox X 4.0.0

Q ADJECTIVE

Summary table

Q ADJECTIVE Hits: 20,181 (79,225.68) Texts: 50/50

HUM

Node

word (lowercase)

Value	Hits	Texts	class	dep	hw	pos	hw_pos*	lemma
other	450	48	1	12	1	1	1	1
such	341	46	1	4	1	1	1	1
social	338	32	1	4	1	1	1	1
different	309	44	1	5	1	1	1	1
more	226	44	1	11	2	1	2	2

Close

Summary table

Q ADJECTIVE Hits: 25,583 (71,416.41) Texts: 61/61

ENG

Node

word (lowercase)

Value	Hits	Texts	class	dep	hw	pos	lemma
high	430	49	1	9	1	1	1
available	417	44	1	12	1	1	1
such	380	54	1	9	1	1	1
different	340	43	1	4	1	1	1
other	336	50	1	10	1	1	1
maximum	244	30	1	8	1	1	1
functional	232	7	1	6	1	1	1
more	230	47	1	9	3	1	3
low	228	41	1	8	1	1	1
large	220	42	1	3	1	1	1
total	218	33	1	7	1	1	1
due	208	44	1	8	1	1	1
current	199	31	1	6	1	1	1
same	196	39	1	9	1	1	1
able	190	38	1	5	1	1	1
constant	189	30	1	16	1	1	1
possible	188	40	1	6	1	1	1
higher	188	44	1	6	1	1	1
initial	181	42	1	5	1	1	1
main	171	39	1	1	1	1	1

Close

Summary table

Q ADJECTIVE Hits: 13,454 (86,020.82) Texts: 50/50

BUS

Node

word (lowercase)

Value	Hits	Texts	class	dep	hw	pos	Lemma
available	308	36	1	11	1	1	1
such	240	44	1	4	1	1	1
social	229	28	1	5	1	1	1
other	197	42	1	8	1	1	1
new	185	40	1	4	1	1	1

Close

Opened KWIC summary table.

Collocation analysis: Collocations

What do you think would be the strongest collocates for the word “social” in the writing of students in the humanities?

whole corpus

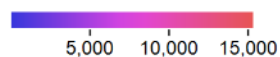
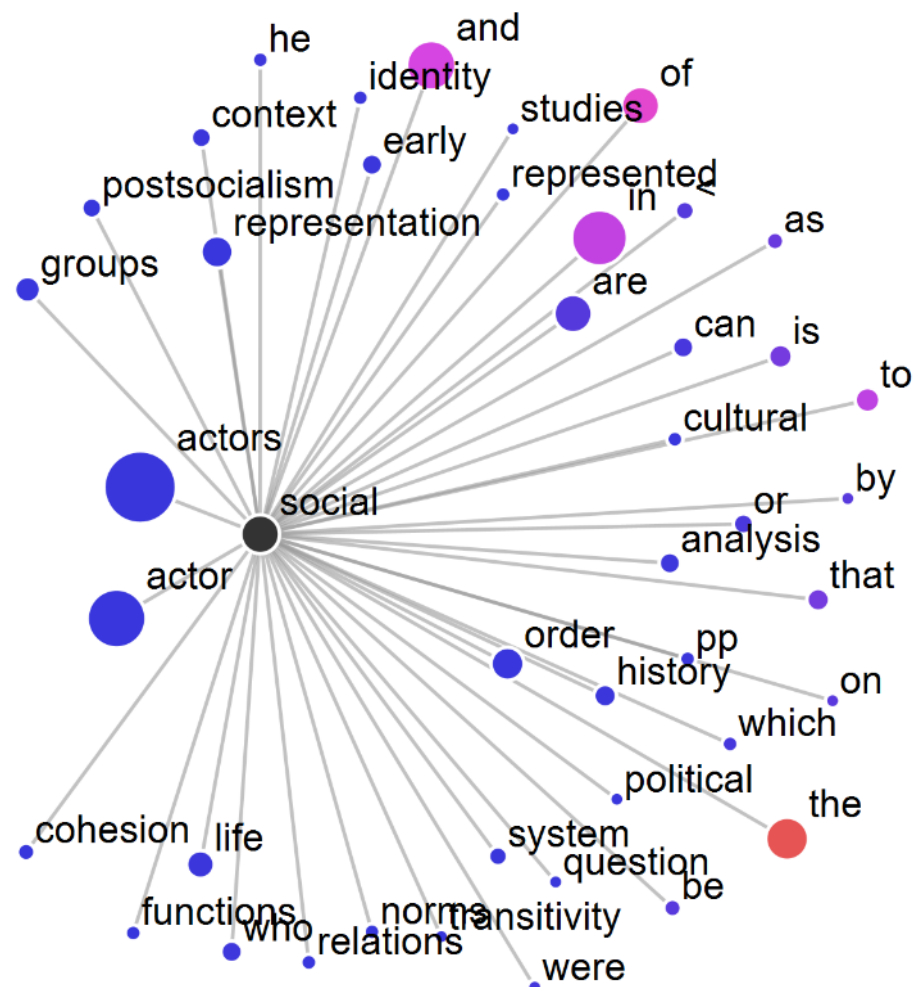
255K

word (lowercase)

R1

R3

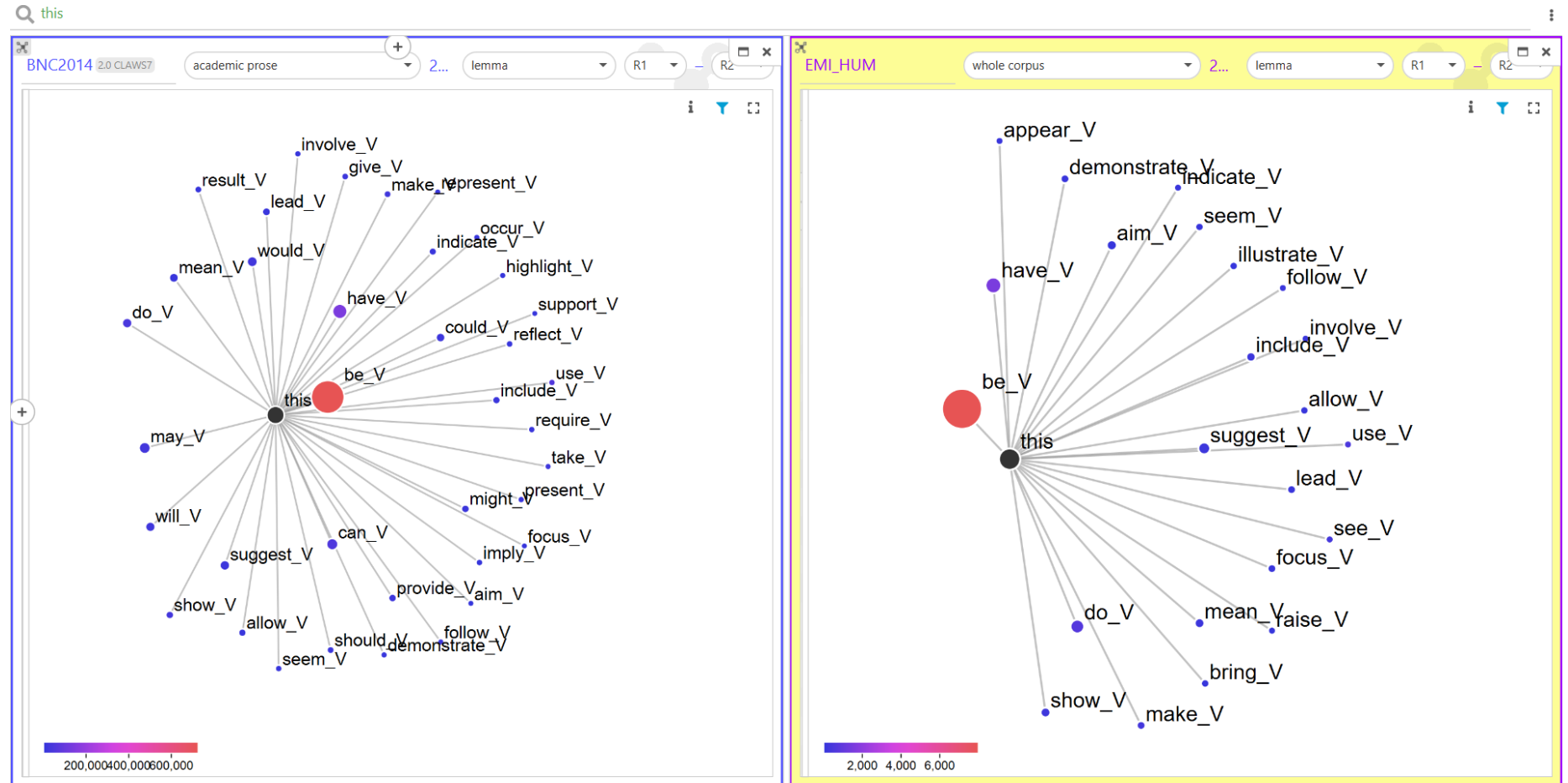
social x			
Q social	Hits: 435 (1,707.70)	Texts: 38/50	Collocates: 482
Y Freq. (collocat			43
Collocate	Freq. (collocat	Freq. (corpus)	Log Dice
functions	6	35	8.7
analysis	10	361	8.7
<	8	1,351	7.2
the	36	15,211	6.2
representation	21	83	10.4
order	22	107	10.4
question	5	123	8.2
were	5	619	7.3
early	10	190	9.0
history	11	230	9.1
cohesion	7	13	9.0
identity	6	184	8.3
cultural	6	106	8.5
system	8	136	8.8
and	47	7,067	7.7
represented	6	62	8.6
are	29	1,370	9.0
relations	6	91	8.5
who	10	351	8.7
as	7	2,177	6.5
be	7	1,467	6.9
by	5	1,392	6.5
actor	66	83	12.0
which	6	705	7.4



ENG			HUM		BUS	
1.	report	129	essay	85	essay	33
2.	project	120	study	49	approach	34
3.	value	75	paper	33	report	21
4.	process	57	test	41	case	15
5.	study	49	case	30	study	13
6.	case	47	way	26	method	13
7.	section	44	period	23	model	15
8.	method	42	section	21	variable	13
9.	model	28	dissertation	16	strategy	14
10.	design	26	approach	17	perspective	12
11.	week	23	research	19	project	18
12.	point	21	time	15	section	10
13.	system	21	point	13		
14.	research	21	image	12		
15.	investigation	20	task	12		
16.	datum	20	listening	11		
17.	equation	19	context	11		
18.	result	19	perspective	10		
19.	tool	17	type	10		
20.	analysis	17	analysis	11		
21.	stage	16	example	11		
22.	time	15				
23.	reaction	15				
24.	problem	15				
25.	simulation	14				

Most frequent nouns
in the 'this + noun'
collocations

Comparing expert vs student writing: 'this + verb' collocations



Created GraphColl tab searching for "this".

Comparing more and less successful student writing

Linguistic category	Example words	Relative frequency per 10,000 words	
		Lower-scored essays	Higher-scored essays
Adjectival expressions of certainty	<i>likely</i>	156.7	196.9
	<i>possible</i>	940.2	492.2
	<i>typical</i>	52.2	69.2
Indefinite pronouns for generalizability	<i>anything</i>	104.5	32.8
	<i>something</i>	156.7	65.5
	<i>everyone</i>	173.4	43.6
Words for hedging	<i>suggest</i>	0	76.1
	<i>might</i>	783.5	63.6
Words for discourse organisation	<i>additionally</i>	52.8	98.4
	<i>thus</i>	64.2	393.8

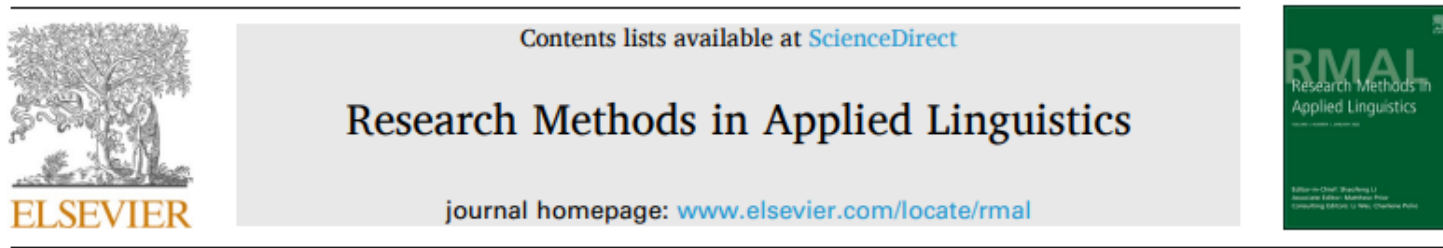
N-gram analysis (lexical bundles, bigrams,...)

What is the most common lexical bundle (3-word expression) starting with 'can' in the writing of students from Engineering?

CAN _____



Exploring EMI corpus for EAP purposes



Gablasova, D., Harding, L., Bottini, R., Brezina, V., Ren, S., Savski, K., Iamartino, G., Li, Y., Liu, T., Poggesi, L., Toomaneejinda, A. & Zottola, A. (2024). Building a corpus of student academic writing in EMI contexts: Challenges in corpus design and data collection across international higher education settings. *Research Methods in Applied Linguistics*, 3(3), 100140.



Challenges in data collection across diverse international education settings

Discussion point

Please discuss the following question with a partner or a small group.

Have you been involved in data collection which involved multiple sites?

What were the benefits of the multi-site data collection?

What were the main challenges that you experienced?



1. A multi-site project: Different educational contexts

- A more **comprehensive picture** of the observed phenomenon
- Increased **ecological validity** of the findings compared to single-site research
- Findings from varied datasets can **inform (pedagogical) practice** across a wider variety of contexts
- **Collaborative nature** of such research and the ability to draw on the collective expertise of team members and their insights into local research sites

Knowledge sharing at different stages of the project:

- **conceptualisation stage** - theoretical frameworks applicable to and inclusive of practices at different research sites;
- **data collection** - enabling collaborators to share experience when issues arise,
- **data analysis and interpretation** - the combined experience and expertise of team members can lead to “a more holistic understanding of findings” (Moranski & Ziegler, 2021, p. 223).

A multi-site project: Challenges

Data collection logs to document challenges and strategies at each individual site

Problem (Aim)
Please use this section to describe and contextualize your issues/aims: 1. What was the aim you were trying to achieve? 2. What were some key/different aspects of this aim? 3. What were the challenges encountered?
Solutions
Please use this section to record the strategies you used, and why they worked or did not work. You may address questions such as: 1. What strategy/strategies have you used? 2. How did the strategy/strategies work for you – why did it work or didn't work? 3. What were the difficult aspects of solving the issue? 4. What helped you with dealing with this challenge?

A multi-site project: Challenges

Getting access to research sites/participants – a potential challenge in any research with human participants - permissions required: **institutional level & level of different academic units** within the institution (e.g. faculty, department)

Multi-site research: permissions differed **in scope and type across institutions** involved in the project – difficult to anticipate/plan for

Example of requirements:

- In some cases, multiple levels of permission required within same institution – e.g. at one research site, an approval was required from the faculty research unit, further approvals from various units within faculty, and an approval from the dean – the same process was repeated for each faculty
- Different practices regarding ethical approval: some institutions accepted LU ethics, others required local ethical approvals

A multi-site project: Challenges

Gaining access – required not only **satisfying the administrative processes** but also required **permission from gatekeepers** (eg. Deans, HoDs, teachers)

The procedure often not completely clear/straightforward

- the request for a permission could take a long time to be considered
- The permission depended not only on administrative procedures but also related to **issues of trust, unfamiliarity with language-related research** and perceived **risks**

Impact on the project: Issues with access affect ability to collect data in some disciplinary areas & some types of assignments

Strategies for addressing challenges in gaining access

1. Being prepared to **communicate the goals** of the project to different audiences -> greater understanding of language-related research led to greater trust and cooperation

Strategies:

- written FAQ documents
- information/discussion sessions for staff in different departments
- recording short videos explaining the project
- showing examples of findings from corpus-based research
- showing examples of previous work completed by the researchers in the team

Strategies for addressing challenges in gaining access

2. Drawing on **existing personal relationships**:

- **for gaining access** to different institutional units (e.g. being able to come to a department to explain what we would like to do)
- shared contacts could help to 'vouchsafe' for the researchers/the project when **establishing new** contacts

3. Prioritising **personal, face-to-face communication**:

- contacting students/departments via emails often led to delays
- personal, face-to-face meetings appeared more effective/efficient in long-term (helping to resolve issues of trust, familiarity with linguistics research, etc)

2. Challenges 'on the ground': Recruiting students

Two major challenges:

- Establishing initial contact
- Gaining consent and obtaining the data

Establishing contact with students & **explaining the project**:

- the need for different context-appropriate strategies
- the strategies differed according to the country, institution, academic unit
- required flexibility and creativity

Challenges 'on the ground': Recruiting students

Strategies: contacting students via departments, using financial incentives (ranging from Amazon vouchers, honoraria, book tokens, coupons for coffee/McDonalds/KFC breakfasts/movies, price draws), contacting students via student reps, social groups; organising information sessions about the project, recording videos and sharing them with students.

While multi-site design made this more challenging – it was also a great source of **solutions** - the combined expertise of the team and understanding of the local context were crucial

- Good understanding of local culture and values
- Sharing ideas about strategies

Discussion point

Please discuss the following questions with a partner/small group:

- As part of your research, did you need to obtain ethics permission for collecting data outside of your home country/institution? If yes, how long did it take to obtain the required permissions?
- Did you experience any challenges when accessing data/participants outside of your home institutions – how did you approach these?



3. Construct of student academic writing

- Decisions about what **language samples** to include in a corpus are central in corpus design → implications for representativeness and generalizability
- Aim of current project – compile a corpus of student writing from different universities and countries → we needed a construct of academic writing that can be **meaningfully applied** across different higher education institutions



Prince of Songkla University



Thammasat University



University of Turin



University of Milan



Xi'an Jiaotong University



Xi'an Jiaotong-Liverpool University

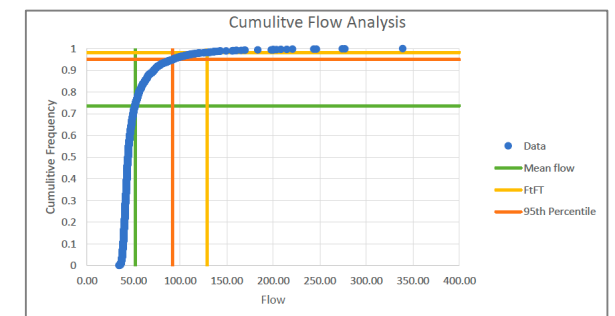
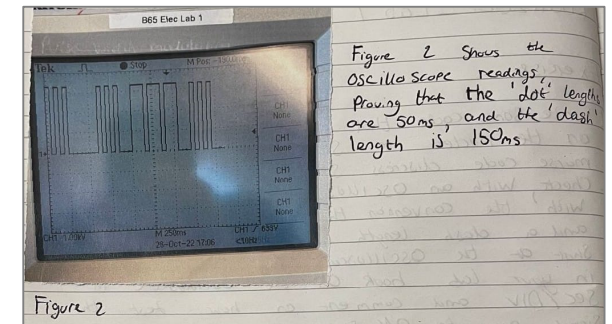


Lancaster University

Operationalising student academic writing: Challenges

- Academic writing is a **complex notion** – can refer to and encompass very **varied set of writing practices** – related to the enormous diversity of academic actors, communicative aims, values, motivations, etc in academic study and research (Hyland, 2006)
- Student writing: *formal assessed pieces – informal notes written during group discussions – emails to course tutors – lecture notes - etc*
- Specific **operationalisation** of the construct → impact on the selection/inclusion of texts → impact on the type of academic writing represented (or excluded) in the corpus

31 An intense rainfall, earthquake shaking, volcanic eruption, storm waves, or rapid stream
32 erosion are causes of increasing the stresses and reducing the strength of slope materials which
33 triggers landslides (Wieczorek 1996). It is also anticipated that incidents of land slide disasters
34 may possibly increase due to over exploitation of natural resources, rapid deforestation, climate
35 change, and increase in hill population and uncontrolled excavations which results in higher
36 susceptibility of surface soil to instability (Manivannan and V. Kasthuri 2020). Van et al. (2010)
37 also add that it is assumed that natural factor are considered as prime factor for the landslide and
38 human activities are considered as less important. Human are regarded as victim of landslide and
39 are considered vulnerable to the disaster but not studied as a factor that might be responsible for



EMI Corpus: Construct of student academic writing

- **All institutions:** Disciplinary writing, submitted for assessment
- Institutions differed in their **preferred types of writing:** Electronic vs handwritten
 - Differences in the type of writing practices and processes (e.g., editing, planning, access to resources, exam setting, effect of stress)

Construct of academic writing: Different writing practices

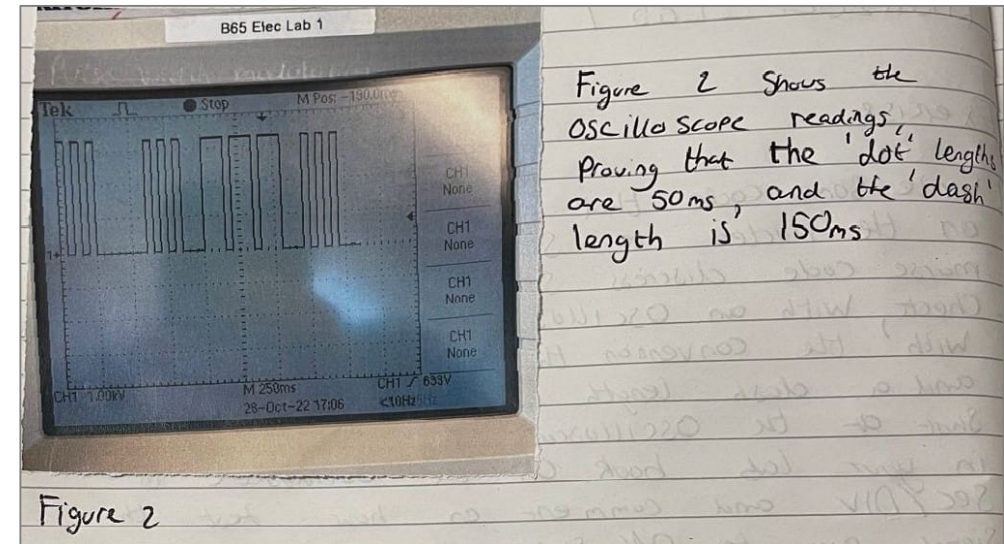
31 An intense rainfall, earthquake shaking, volcanic eruption, storm waves, or rapid stream
32 erosion are causes of increasing the stresses and reducing the strength of slope materials which
33 triggers landslides (Wieczorek 1996). It is also anticipated that incidents of land slide disasters
34 may possibly increase due to over exploitation of natural resources, rapid deforestation, climate
35 change, and increase in hill population and uncontrolled excavations which results in higher
36 susceptibility of surface soil to instability (Manivannan and V. Kasthuri 2020). Van et al. (2010)
37 also add that it is assumed that natural factor are considered as prime factor for the landslide and
38 human activities are considered as less important. Human are regarded as victim of landslide and
39 are considered vulnerable to the disaster but not studied as a factor that might be responsible for

Answers to the question no :- 3

Innovation is the process of creating a new product, process, organizational methods or marketing methods in a new or improved way of that existing identity product, process, organizational methods or marketing

Writing practices reflecting different contexts of production → typical linguistic features

- Handwritten vs electronically submitted
- Produced in timed vs non-timed conditions
- Produced under exam conditions

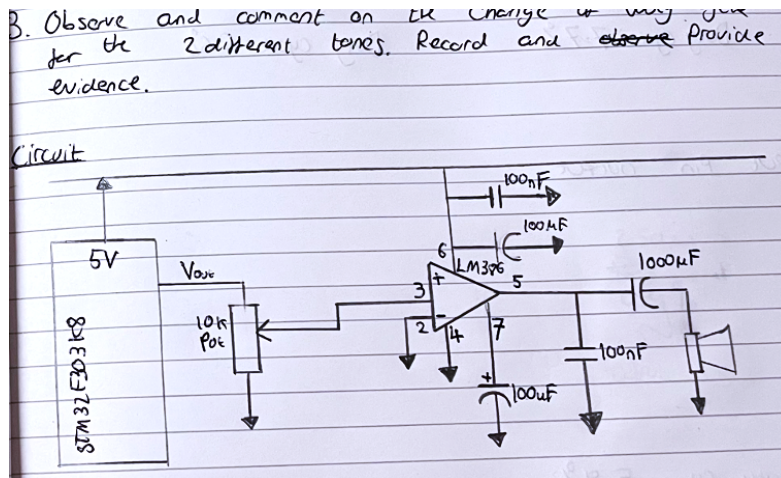
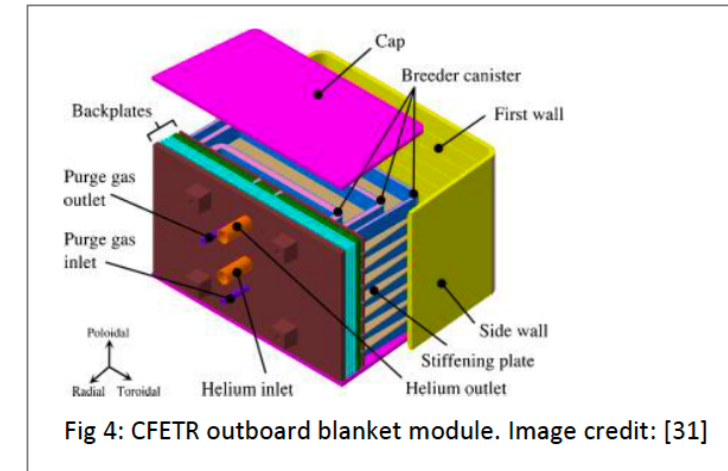
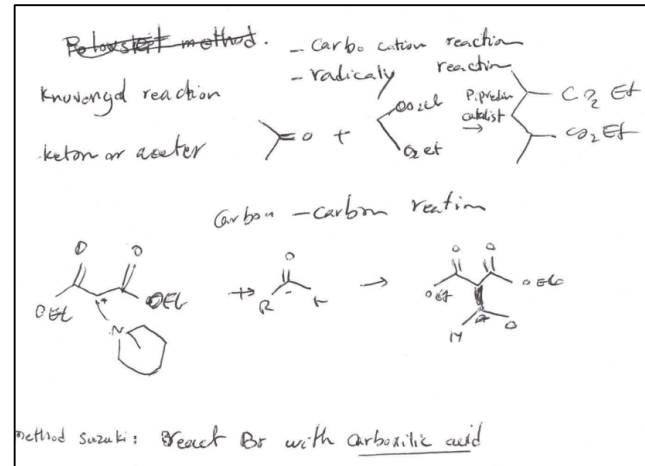


EMI Corpus: Construct of student academic writing

- Disciplinary writing, submitted for assessment
- Electronic & handwritten submissions
- **Word length** varied considerably across **disciplines and institutions**: Written pieces – min. 100 words - including text, figures, diagrams, code, etc.

Construct of academic writing: Different writing practices

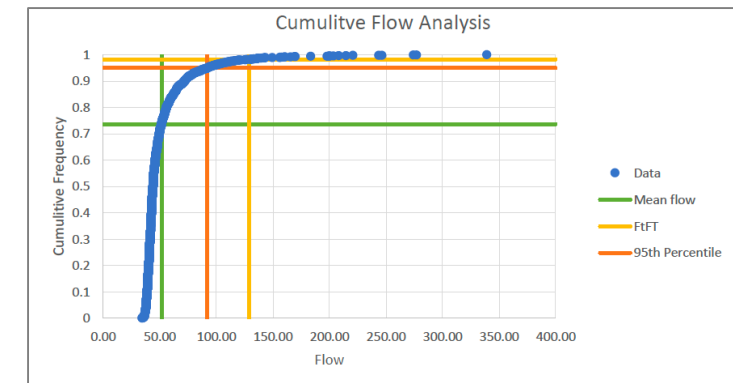
Capturing the **visual aspects** of student production → insights into the changing nature of what counts as 'academic writing' and in what way this differs across disciplines (e.g. STEM subjects)



residence time is calculated below:

$$T_{B,mn} = \left(\frac{\mu_{mx} C_{P,i}}{k_{sat} C_{P,i}} - k_D \right)^{-1}$$

$$T_{B,mn} = \left(\frac{(0.023)(107.07)}{(47.58)(107.07)} - (0.0013) \right)^{-1} = 70.00 \text{ hr}$$



EMI Corpus: Construct of student academic writing

- Disciplinary writing, submitted for assessment
- Electronic & handwritten submissions
- Written pieces – min. 100 words - including text, figures, diagrams, code

Construct of student academic writing

The adopted construct prioritised – as much as possible – an **inclusive approach** – to maximise the opportunities offered by access to multiple educational sites

The final approach followed close discussions with partner universities to understand:

- i) the **type of writing** produced but also ii) the **status** of different types of writing and
- iii) the local **writing practice/needs**

Discussion point

Addressing challenges related to deciding on the construct of 'academic writing' in research/teaching

Please discuss the following questions with a partner or a small group:

- What type of student academic writing would you like to see represented in research/corpora?
- Are there particular challenges when collecting the target data across different institutions?



Concluding thoughts

The session highlighted...

- some of the **challenges** involved in a multi-site, international project involving corpus construction and the strategies/ approaches used to address them
- the **interaction of theoretical, methodological and practical** considerations that are part of collaborative projects that involve data collection at different sites
- the value of (large) **collaborative projects** that can draw on the **experience and expertise** of researcher from different educational backgrounds and settings

Thank you!

Corpus research informing EMI practice



EMI CORPUS PROJECT

Future of English research programme

<https://wp.lancs.ac.uk/emi-corpus-project/>