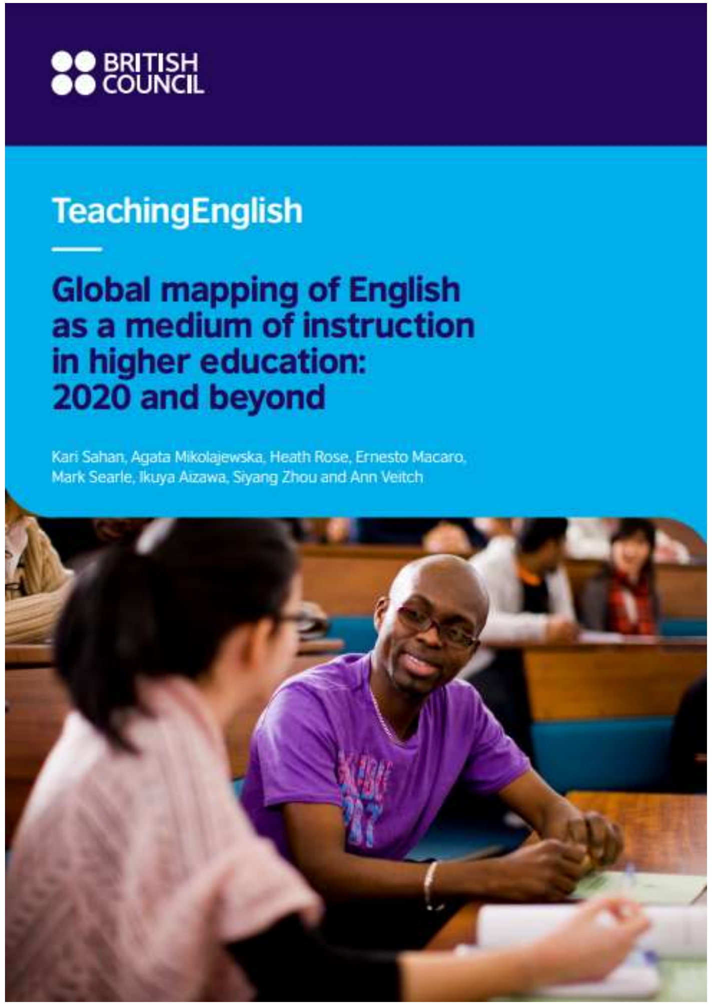
Building a corpus of student academic writing in EMI contexts: *Challenges in data collection across international higher education settings*

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

Lancaster University

English medium instruction (EMI)



BRITISH COUNCIL

TeachingEnglish

Global mapping of English as a medium of instruction in higher education: 2020 and beyond

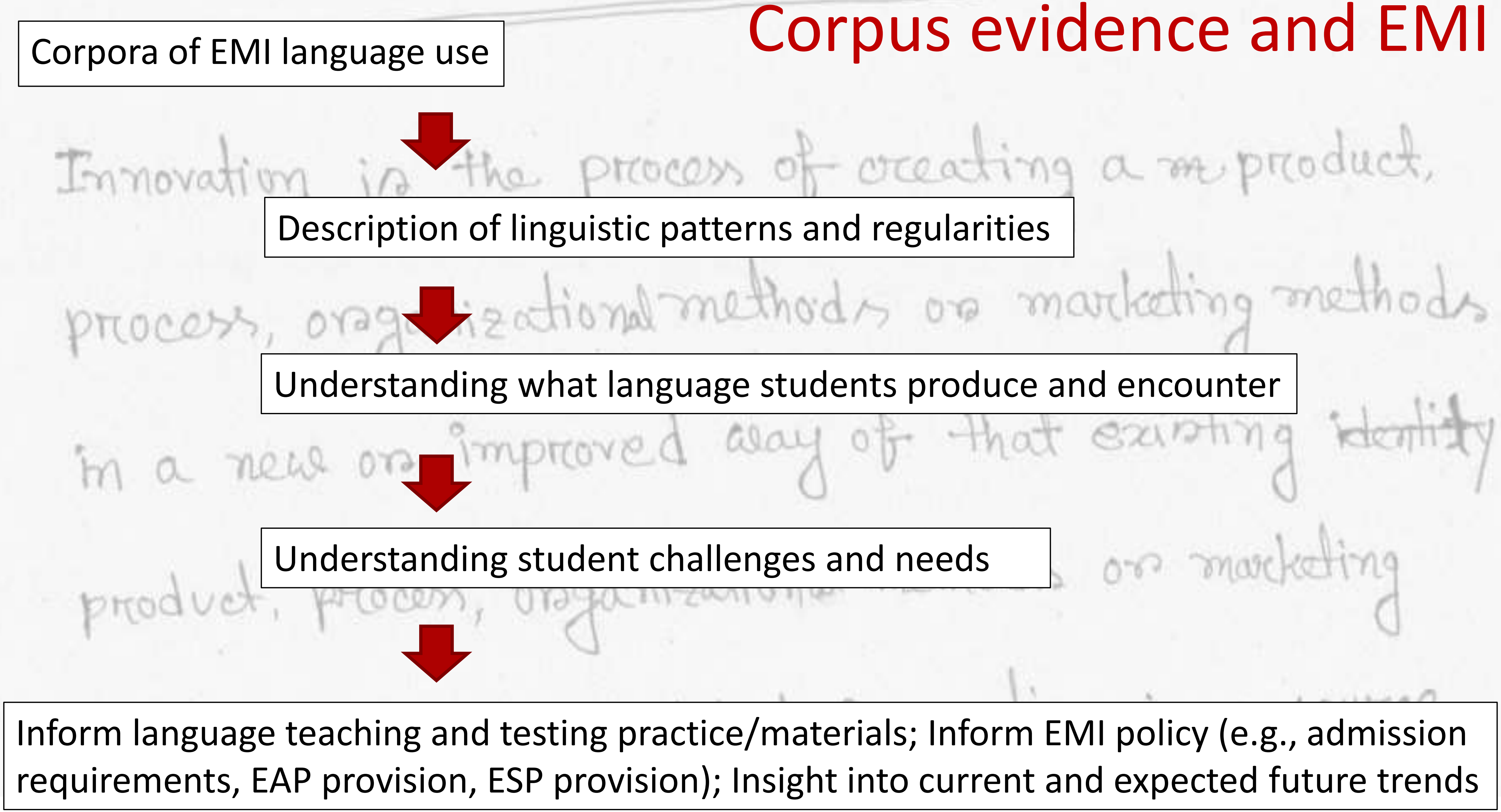
Kari Sahan, Agata Mikolajewska, Heath Rose, Ernesto Macaro, Mark Searle, Ikuya Aizawa, Siyang Zhou and Ann Veitch

- EMI – teaching/learning disciplinary subjects through the medium of English in the countries where English is not the community language
- EMI – currently a global pedagogical trend; on the increase
- EMI advantages: individual, national, international
- Use and knowledge of English – crucial for understanding subject knowledge and for learning

EMI: Challenges

- 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)

Corpus evidence and EMI



Corpora of EMI language use

Description of linguistic patterns and regularities

Understanding what language students produce and encounter

Understanding student challenges and needs

Inform language teaching and testing practice/materials; Inform EMI policy (e.g., admission requirements, EAP provision, ESP provision); Insight into current and expected future trends

Answer to the question no :- 2

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

product, process, organizationa[illegible] or marketing

Corpus research informing EMI practice

EMI Corpus project

Lancaster University

BRITISH COUNCIL

Project: "Linguistic demands of EMI in Higher Education: A corpus-based analysis of student writing and reading in EMI university settings in China, Italy, Thailand and the UK"

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

Aims:

- Contribute to the description of EMI across different higher educational contexts (countries/institutions)
- Contribute to the existing datasets (e.g. BAWE, MICUSP) available for a systematic research on student English writing at university level
- Inform language teaching/testing related to EMI (e.g., admission requirements, teaching resources)

Prince of Songkla University

Thammasat University

University of Turin

University of Milan

Xi'an Jiaotong University

Xi'an Jiaotong-Liverpool University

EMI Corpus: Data collection (2022-2024)

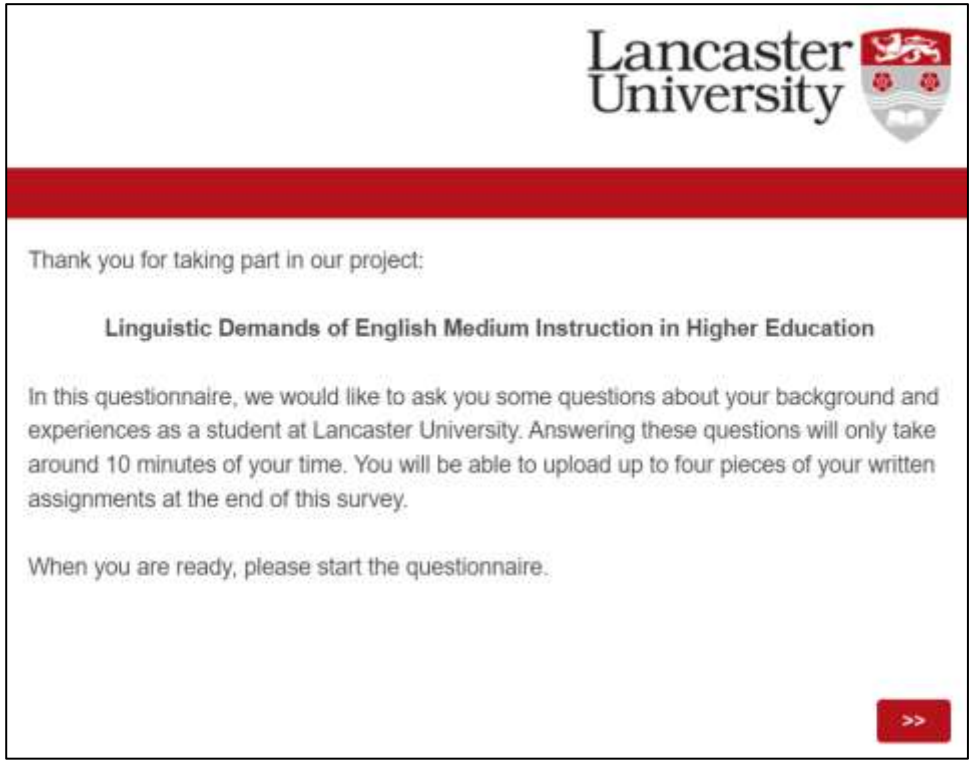
Types of data collected:

Student background data

- Demographic information (e.g. age, L1, proficiency)
- Academic reading/writing habits

Student writing

- Written texts
- Information about the assignments (e.g. mark, instructions)



Thank you for taking part in our project:

Linguistic Demands of English Medium Instruction in Higher Education

In this questionnaire, we would like to ask you some questions about your background and experiences as a student at Lancaster University. Answering these questions will only take around 10 minutes of your time. You will be able to upload up to four pieces of your written assignments at the end of this survey.

When you are ready, please start the questionnaire.

Data collection framework

Level	**UG**	**PG**			
Disciplinary area	Business & Management	Business & Management	Humanities & Social Science	Life sciences	Science & technology
Core subjects	Business studies, Economics, Management, Finance, Accounting, Administration		History, Literature, Sociology, Linguistics	Chemistry, Biology	Engineering, Computer science
Balance	20%	20%	20%	20%	20%

Current corpus size: 3M words from over 1,000 student texts

Our team (Lancaster University)

Dana Gablasova (PI)

Luke Harding

Vaclav Brezina

Raffaella Bottini

Haoshan Ren

Our research partners

Dr. Kristof Savski
Prince of Songkla University

Dr. Angela Zottola
University of Turin

Dr. Yingyu Li
Xi’an Jiaotong University

Dr. Anuchit Toomaneejinda
Thammasat University

Prof. Giovanni Iamartino
University of Milan

Dr. Tanjun Liu
Xi’an Jiaotong-Liverpool University

EMI corpus: Challenges in corpus design and data collection across international higher education settings

Construct of student academic writing

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 need 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

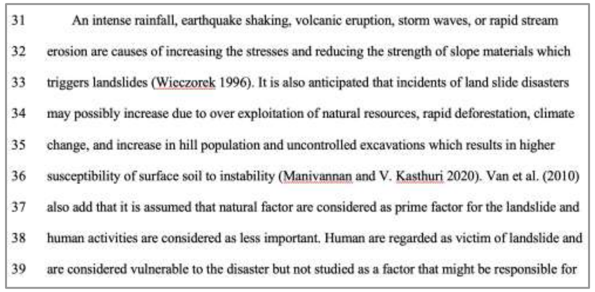
Lancaster University

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



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

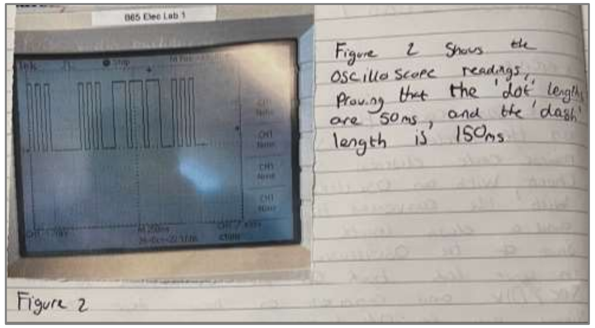
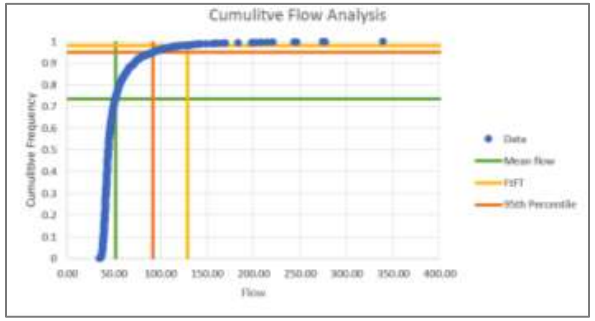


Figure 2 Shows the oscillo scope readings, Proving that the 'dot' lengths are 50ms, and the 'dash' length is 150ms

Figure 2

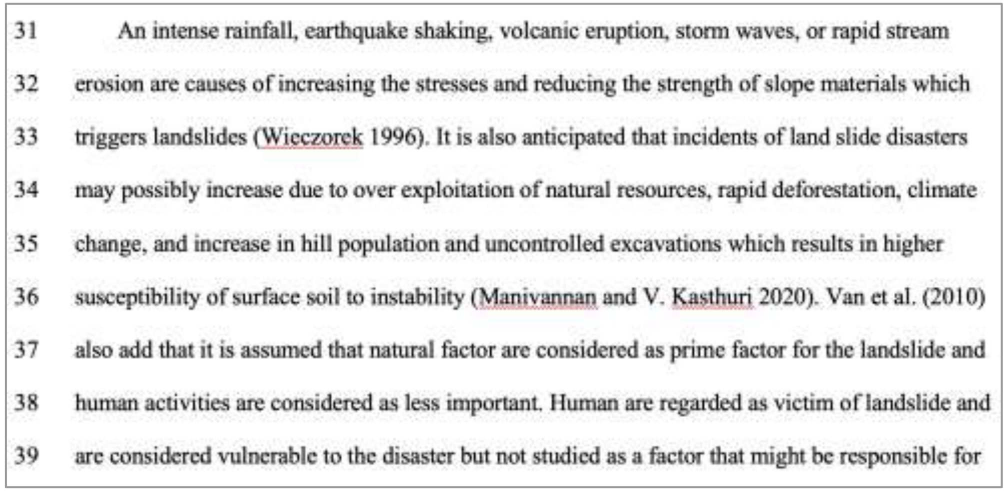


EMI Corpus: Construct of student academic writing

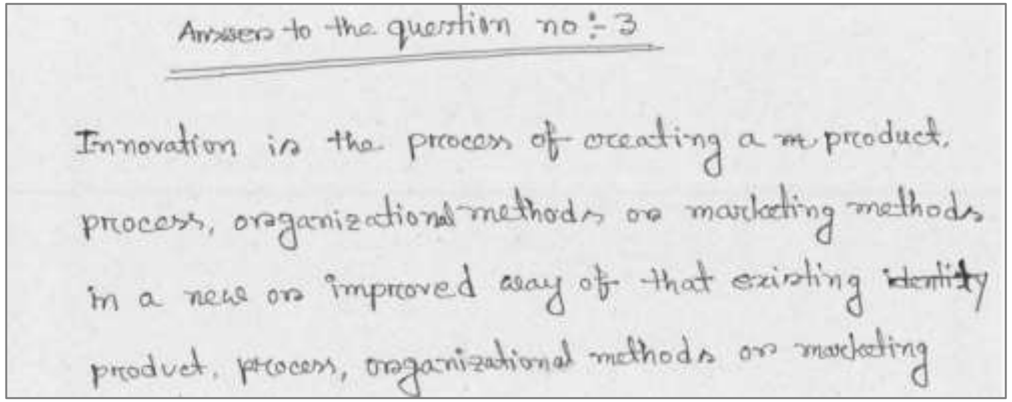
- Disciplinary writing, submitted for assessment

- Electronic & handwritten submissions
  - Differences in the type of writing practices and processs (e.g., editing, planning, access to resources, exam setting, effect of stress)

Construct of academic writing: Different writing practices



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



Answer to the question no:- 3

Innovation is the process of creating a ~~m~~ 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

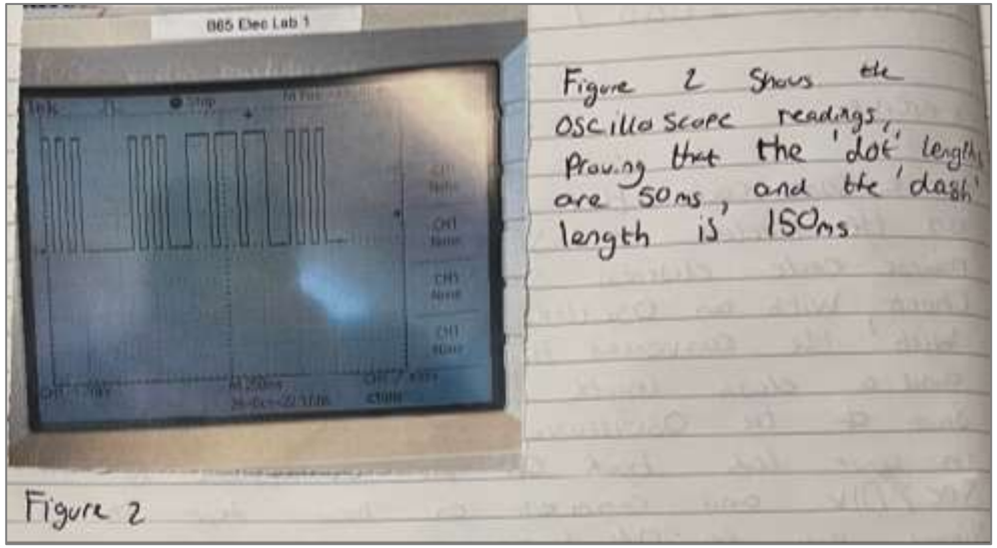


Figure 2 Shows the oscillo Scope readings, Proving that the 'dot' length are 50ms, and the 'dash' length is 150ms

Figure 2

EMI Corpus: Construct of student academic writing

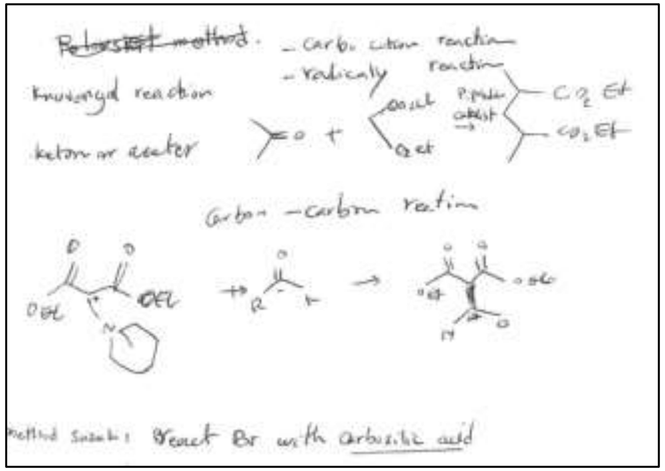
- Disciplinary writing, submitted for assessment

- Electronic & handwritten submissions

- Written pieces – min. 100 words - including text, figures, diagrams, code, 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)



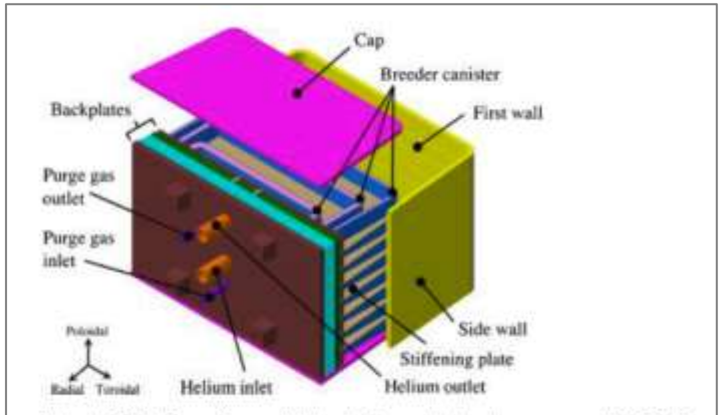
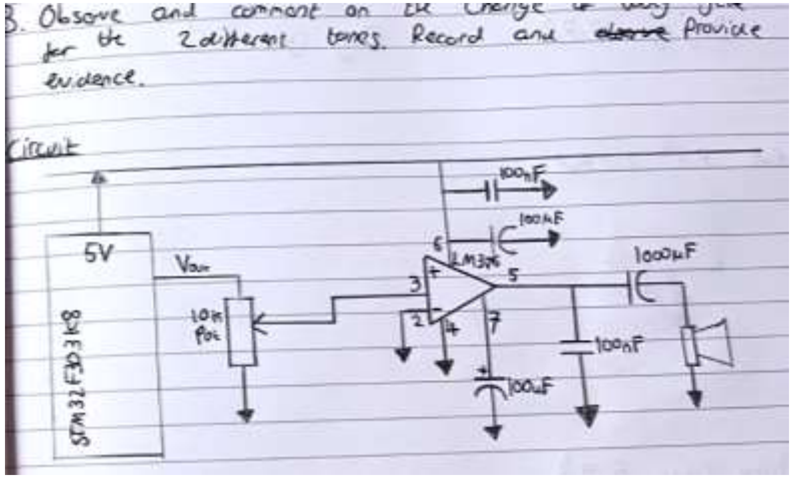
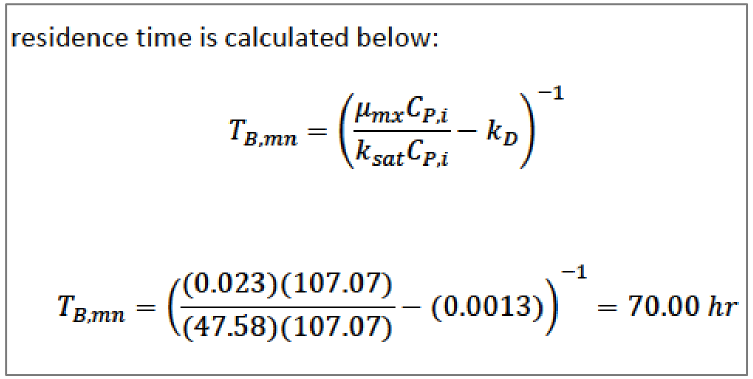


Fig 4: CFETR outboard blanket module. Image credit: [31]



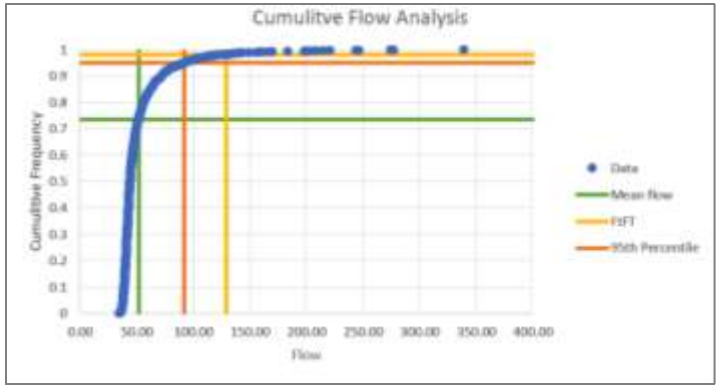
3. Obsorve and comment on the change [illegible] for the 2 different tones. Record and ~~above~~ Provide evidence.



residence time is calculated below:

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

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



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
- Quality: pass
  - Satisfies the requirements for a passable university standard; but the mark is recorded so possible to distinguish higher/lower marks; issue of dealing with different marking systems (e.g. what a 'pass' is across universities)

Construct of student academic writing

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

Adopting broader criteria:

- Theoretical implications: capturing the complexity and variation in EMI writing
- Methodological implications: greater 'messiness' of the data and greater challenges for data processing (e.g. dealing with equations, digitising hand-written texts)
- Practical implications: project feasibility → higher demands on time and resources

Different educational contexts:

A multi-site transnational project

A multi-site project: Benefits

Comprehensive insights into observed phenomenon:

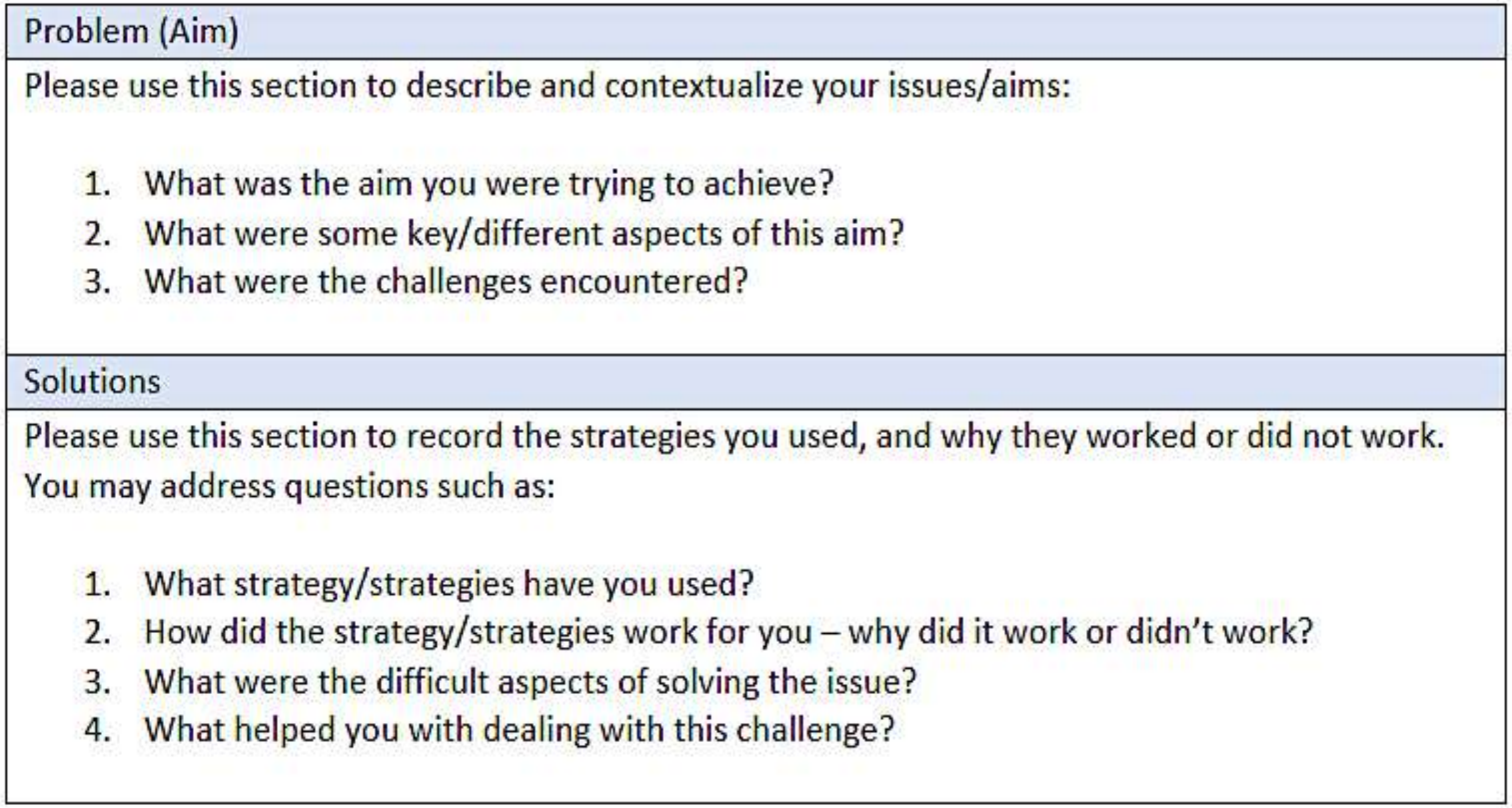
- Enhances representativeness and diversity of data
- Increases ecological validity of the findings
- Can inform (pedagogical) practice across a wider variety of contexts
- Offers ability to draw on the collective expertise of team members and their insights into local research sites (Kwon et al, 2018).

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?

Gaining access across institutional barriers

- Getting access to research sites/participants – a potential challenge in any research with human participants

- Two dimensions:
  - Addressing institutions and their administrative requirements
  - Working with institutional gatekeepers

- Both dimensions were crucial in the EMI Corpus project

Addressing institutional administrative requirements

- 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

Working with institutional gatekeepers

- 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**

Strategies for institutional challenges

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 institutional challenges

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)

Recruiting students: Challenges

Two major challenges have been involved:

- 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

Recruiting students: Strategies

Strategies: contacting students via departments, using financial incentives in an effective way (e.g. ranging from Amazon vouchers, honoraria, book tokens, coupons for coffee/McDonalds/KFC breakfasts/movies, price draws, etc), 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 for solutions:

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

Summary

- We highlighted some of the **challenges** involved in a multi-site, international corpus construction process and the strategies/approaches used to address them

- It is important to **reflect on and record** the challenges and decision-making process in corpus development
  - The users can understand better the data and type of evidence in the corpus
  - To highlight the interaction of theoretical, methodological and practical considerations that are part of creating a new dataset

Thank you!