Structured by ADDIE, Guided by Freire:

Developing Critical MT Quality Assessment Training

João Lucas Cavalheiro Camargo

Lic, M.Ed.

A dissertation submitted in fulfilment of the requirements for the award of Doctor of Philosophy (PhD)

to the

Dublin City University

School of Applied Language and Intercultural Studies

Supervisors:

Dr. Sheila Castilho

Dr. Joss Moorkens

July 2026

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of PhD is entirely my own work, and that I have exercised reasonable care to ensure that the work is original and have conformed to the regulations on the use and declaration of Generative AI, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

João Lucas Cavalheiro Camargo

ID No.: **21267996**

Date: July 13, 2026

List of Generative AI Tools

Tool	Output	Prompts & Iterations	Critical Reflection
Google Translate	Translation of texts for the compilation of a pedagogical dataset	This GenAI tool required no prompt and only allowed the generation of output through a button "Translate" or automatically in the UI.	Machine Translation is the main GenAI task of this PhD thesis.
DeepL	Translation of texts for the compilation of a pedagogical dataset	This GenAI tool required no prompt and only allowed the generation of output through a button "Translate" or automatically in the UI.	Machine Translation is the main GenAI task of this PhD thesis.
Amazon Translate	Translation of texts for the compilation of a pedagogical dataset	This GenAI tool required no prompt and only allowed the generation of output through a button "Translate" or automatically in the UI.	Machine Translation is the main GenAI task of this PhD thesis.
ChatGPT -3	Text Generation, Summarisation, Text Structuring for paraphrasing of sentences and paragraphs to be edited and adapted for original contribution; Grammatical correction; Formatting for OverLeaf/LaTeX	The date of use and prompts were not documented because, at the time, systematic prompt recording was not an established element of the methodological protocol, and iterations were conducted in a manner designed to preserve the originality of the model outputs and maintain compliance with the project's data-handling and cybersecurity requirements. Estimated date of use - 30/11/2022 to 14/03/2023.	This tool was used with the intention of exploring its applicability to academic practices, particularly for automating routine tasks in order to increase the efficiency and speed of the work. Every output was checked for its factuality in academic literature and/or search engines leading to reputable sources. Every output was generated with the intention of editing to be performed (light or heavy) with the accordance of my own expertise and needs.

Tool	Output	Prompts & Iterations	Critical Reflection
ChatGPT -3.5	Text Generation, Summarisation, Text Structuring for paraphrasing of sentences and paragraphs to be edited and adapted for original contribution; Grammatical correction; Formatting for OverLeaf/LaTeX	The date of use and prompts were not documented because, at the time, systematic prompt recording was not an established element of the methodological protocol, and iterations were conducted in a manner designed to preserve the originality of the model outputs and maintain compliance with the project's data-handling and cybersecurity requirements. Estimated date of use - 30/11/2022 to 14/03/2023.	This tool was used with the intention of exploring its applicability to academic practices, particularly for automating routine tasks in order to increase the efficiency and speed of the work. Every output was checked for its factuality in academic literature and/or search engines leading to reputable sources. Every output was generated with the intention of editing to be performed (light or heavy) with the accordance of my own expertise and needs.
ChatGPT -4	Text Generation, Summarisation, Image Generation, Text Structuring, paraphrasing of sentences and paragraphs to be edited and adapted for original contribution; Generation of images for pedagogical use; Formatting for OverLeaf/LaTeX	The date of use and prompts were not recorded because prompt logging was not yet part of the methodological workflow, and interactions were frequently conducted in Temporary Mode as a good privacy practice to avoid exposing any data or original work to a third-party service that might use such information for its own purposes. Estimated date of use - 14/03/2023 to 13/05/2023.	This tool was used with the intention of exploring its applicability to academic practices, particularly for automating routine tasks in order to increase the efficiency and speed of the work. Image Generation was used to create custom characters that fit the diversity needs of ready-to-use images to include stylised cartoon images of non-binary individuals, people of colour, etc. to be added to educational materials. Every output was checked for its factuality in academic literature and/or search engines leading to reputable sources. Every output was generated with the intention of editing to be performed (light or heavy) with the accordance of my own expertise and needs.

Tool	Output	Prompts & Iterations	Critical Reflection
ChatGPT -4o	Text Generation, Summarisation, Image Generation, Text Structuring, paraphrasing of sentences and paragraphs to be edited and adapted for original contribution; Generation of images for pedagogical use; Formatting for OverLeaf/LaTeX	The date of use and prompts were not recorded because prompt logging was not yet part of the methodological workflow, and interactions were frequently conducted in Temporary Mode as a good privacy practice to avoid exposing any data or original work to a third-party service that might use such information for its own purposes. Estimated date of use - 13/05/2024 to 07/08/2025.	This tool was used with the intention of exploring its applicability to academic practices, particularly for automating routine tasks in order to increase the efficiency and speed of the work. It was used to generate initial examples or templates of pedagogical activities that were edited according to my own pedagogical expertise and instructional design requirements. Image Generation was used to create custom characters that fit the diversity needs of ready-to-use images to include stylised cartoon images of non-binary individuals, people of colour, etc. to be added to educational materials. Every output was checked for its factuality in academic literature and/or search engines leading to reputable sources. Every output was generated with the intention of editing to be performed (light or heavy) with the accordance of my own expertise and needs.

Tool	Output	Prompts & Iterations	Critical Reflection
ChatGPT -5	Text Generation, Summarisation, Image Generation, Text Structuring for paraphrasing of sentences and paragraphs to be edited and adapted for original contribution; Grammatical correction; Generation of graphics from text; Formatting for OverLeaf/LaTeX	The date of use and prompts were not recorded because prompt logging was not yet part of the methodological workflow, and interactions were frequently conducted in Temporary Mode as a good privacy practice to avoid exposing any data or original work to a third-party service that might use such information for its own purposes. The "Personalised Prompt" of this tool, which applied to every output generated, was "Always estimate the energy consumption of my AI interactions, including text and image generation. For EVERY SINGLE PROMPT. Adjust estimates based on prompt complexity. If I generate multiple responses or images, keep a running tally. Periodically remind me of my estimated cumulative energy consumption. Always use British English unless requested otherwise. If you cannot provide information, instead of making up, please tell me outright you can't do it. Do not try to please me, be as technical as possible because I am an NLP researcher." Estimated date of use - 13/05/2024 to 24/11/2025.	This tool was used with the intention of automating routine tasks in order to increase the efficiency and speed of the work. Every output was checked for its factuality in academic literature and/or search engines leading to reputable sources. Every output was generated with the intention of editing to be performed (light or heavy) with the accordance of my own expertise and needs.

Tool	Output	Prompts & Iterations	Critical Reflection
Writefull	Text Generation for paraphrasing of sentences and paragraphs to be edited and adapted for original contribution/edition; Grammatical correction	There are no prompts in this tool, merely suggestions or the option to highlight pieces of text to request for paraphrasing. Estimated date of use - 19/01/2024 to 24/11/2025.	This tool was used with the intention of automating routine tasks in order to increase the efficiency and speed of the work. Every output was checked for its factuality in academic literature and/or search engines leading to reputable sources. Every output was generated with the intention of editing to be performed (light or heavy) with the accordance of my own expertise and needs.
Canva	Image Enhancement, Image Editing for Upscaling of images and graphics; Removal or addition of elements in images	There are no prompts for this type of tool. I clicked menu options such as "Background remover", "Image Enhancement". Estimated date of use - 01/01/2023 to 24/11/2025	This tool was employed to improve the clarity of upscaled images, a key consideration given my visual disability. Clear text within images is essential both for my own accessibility needs and for ensuring that the educational materials I produce meet appropriate pedagogical accessibility standards.

Acknowledgements

Firstly, I would like to express my gratitude to both my supervisors, Dr. Sheila Castilho and Dr. Joss Moorkens. For Dr. Castilho, I am extremely thankful for the patience, expertise, countless examples, resources and preciseness in feedback. For Dr. Moorkens, I offer extreme appreciation towards your meticulous feedbacks that intertwined translation theory, methodology and practicalities in such a clear way. You two composed the best duo of supervisors that I could have ever asked for. I appreciate you for never giving up on me, even if at many times I gave you reasons to do so. I truly appreciate your roles as mentors, as I would not be able to finish this PhD without you.

I am extremely grateful to the School of Applied Language and Intercultural Studies for providing me with a scholarship to carry out this PhD, as without it I would never have been able to accomplish it. I am also very grateful to all staff from the Faculty of Humanities and Social Sciences, as many professors and colleagues provided me with many insights, feedback, ideas, and opportunities to collaborate and work together.

I would also like to thank the ADAPT Centre for the office space and many funding opportunities, events, and training that otherwise would not have been made available to me.

I also would like to thank everyone who contributed to my fundraiser done in 2021, as I did not have the resources to move to Ireland. Without your help, I would not have been able to buy the plane tickets or get on with other practicalities that this move entails. People like me typically do not get to do PhDs, so, without your help, I would not have been here writing this thesis right now. I recognise that many of those who contributed were friends, colleagues and former professors of mine. Muito obrigado.

I thank my great colleagues on my first years of the PhD, Vicent Briva-Iglesias and Matt Riemland. It was good craic to go on conferences, talk about translation research, and learn from your experiences. We even managed to collaborate a few times, which was truly an amazing opportunity. In the last years of my PhD, I want to thank my friends from the ADAPT Centre, who adopted me once I moved to the computing side of research. Michela, Charlie, Kanishk, Tytti, and other colleagues from the lab, you always explained to me with the utmost patience my questions about Natural Language Processing and Computer Science in general, so I will forever be thankful. And thanks to all my friends in DCU who have impacted my journey in one way or another.

To my friends from DCU International Academy, who provided me the chance to have a memorable opportunity to work alongside you. It was lovely working part-time over the past four years, much of what I have learned about Ireland, about my newly found Dublin accent and about the workplace here, I have learned from you.

Thank you very much Stephen and Cathal for the opportunities you have given to me, I will never forget your kindness. And thank you to Rachel, Amy, Agata, Aoife, Rita, Janet for being extremely fun coworkers.

To Jamie Kenny, although we walk different paths now, I still care about you very much and you were incredibly helpful and meaningful in my journey. You know...two laptops, an energetic cocker spaniel, video games and countless clothes that I took from you to wear on conferences, networking events and teaching. As Buffy once told Angel: "Because, okay...I am cookie dough, and I am not done baking". My gratitude will always be there. Thank you, my good friend.

To my special friends who made Ireland feel like home, I have found family in you. Thank you for checking in, inviting me out for drinks, and being there for me when I needed. David and Chris, you were there for me at my absolute worst and offered all the support you could, I will forever hold our friendship deep in my heart with the utmost care. Jonny and Par, you (and your lovely dogs) were also there for me when I needed the most help, I send the warmest of hugs. Tom, Chris, Cathal, Joe, Germán, Neill, and Jameson, thank you for the fun adventures, non-sensical conversations, and frequent trips to Street 66. Billie and Violet, thank you for the lovely evenings in Love Tempo, where we tried to have fun conversations while we all stopped Buffy from eating the wall. I love you, friends.

To my family and friends in Brazil, I will thank you in the most appropriate way. Muito obrigado por tudo. Para minha mãe Gisele, meu pai João e meu irmão José, muito obrigado por terem apoiado. Eu sei que não pude estar com vocês em momentos que mais precisavam, por isso, eu peço desculpas. Minhas saudades sempre são eternas. Eu finalmente conquistei um dos meus sonhos, e eu espero que isso possa dar orgulho a vocês, mesmo que eu não seja o melhor filho ou o melhor irmão. Aos meus amigos do Brasil, que estiveram comigo por mensagens de voz ou visitas infrequentes, obrigado pela companhia. Aos meus professores e colegas da Universidade Estadual do Oeste do Paraná, os aprendizados que ganhei ao estudar com vocês e eventualmente trabalhar ao seu lado, seus ensinamentos me ajudaram profundamente no meu doutorado.

This PhD was quite the ride. The João who stands here writing these acknowledgments is certainly not the same João who first arrived in Ireland to do his PhD. Many things changed for me, my dreams, my desires, my goals. I grew tremendously through this experience. Growth will always define this stage of my life. This PhD is a forever reminder that I am an imperfect human but one who desires to learn and improve. May this PhD thesis be useful in the future to an educator who wants to change this world, especially for those who come after.

Contents

List of Generative AI Tools iii

Contents x

List of Figures xvi

List of Tables xxi

List of Abbreviations xxiv

1 Introduction 1
- 1.1 Context . . . 1
- 1.2 Motivation . . . 5
- 1.3 Research Questions . . . 6
- 1.4 Thesis Structure . . . 10

2 Epistemological Rationale 12
- 2.1 Pedagogical Rationale . . . 12
- 2.2 Theoretical Assumptions and Methods . . . 13
  - 2.2.1 Foundations of Translation . . . 14
    - 2.2.1.1 The Goals of Machine Translation . . . 15
    - 2.2.1.2 The Epistemology of Translation and its Assessment 15
    - 2.2.1.3 Translation and evaluation as interpretation, translators and evaluators as active entities . . . 17
    - 2.2.1.4 Interpretation as a mediation of values, expertise and ethics . . . 20
    - 2.2.1.5 Translation Hermeneutics and Evaluation . . . 21
    - 2.2.1.6 Translation Hermeneutics and other translation theories . . . 24
    - 2.2.1.7 Translation Hermeneutics and relations to human or machine translation . . . 25
    - 2.2.1.8 Machine Translation Quality Assessment . . . 28
  - 2.2.2 Foundations of Education and Instructional Design . . . 31
    - 2.2.2.1 From epistemology to curriculum and practice - Freirean Philosophy . . . 32
    - 2.2.2.2 Instructional Design . . . 34

3 Systematic Review of Literature - "Needs Analysis" 37
- 3.1 What is a systematic review of literature? . . . 38

3.2 Systematic Review of Literature Protocol . . . 39
3.3 Data Collection Methods . . . 42
3.4 Data Coding . . . 44
3.5 Data Synthesis . . . 46
3.5.1 Stage 1 - Translation Studies . . . 46
3.5.1.1 Rationale (RTL) . . . 49
3.5.1.2 Resources (RE) . . . 53
3.5.1.3 Modality (MOD) . . . 56
3.5.1.4 Teaching Procedures and Methods (TPM) . . . 58
3.5.1.5 Curriculum and Syllabus (CSL) . . . 60
3.5.2 Stage 2 - Natural Language Processing . . . 63
3.5.2.1 Curriculum and Syllabus (CSL) . . . 63
3.6 Principles towards HE in NLP education . . . 65
3.6.1 Principle 1: Emphasis on Human-Centric Evaluation in MTQA Training . . . 66
3.6.2 Principle 2: Adaptable and Contextualised MTQA Training . 66
3.6.3 Principle 3: Integrative Evaluation Platforms and Resources . 67
3.6.4 Principle 4: Emphasis on Reflective Practice in the classroom 67
3.7 Curricular Analysis - MT in the NLP curriculum . . . 68
3.7.1 Inclusion and Exclusion Criteria . . . 69
3.7.2 Results . . . 70
3.7.2.1 Programmes . . . 70
3.7.2.2 Modules . . . 71
3.7.2.3 Metrics . . . 72
3.7.3 Discussion . . . 74
3.8 Conclusion . . . 75

4 Methods - "Design" 78
4.1 Phase 1c - Survey With Teachers . . . 79
4.1.1 Survey Setup . . . 80
4.1.1.1 Participant Profile . . . 80
4.1.1.2 The Research Instrument - Questionnaire . . . 80
4.1.2 Data Analysis . . . 83
4.1.2.1 Block 1 - Teacher Profile . . . 83
4.1.2.2 Block 2 - Human Evaluation Predictions . . . 88
4.1.2.3 Block 3 - MTQA Pedagogical Questions . . . 90
4.1.2.4 Block 4 - Human Evaluation Teaching Recommendations . . . 95
4.2 Phase 2 - Course Blueprints . . . 97
4.2.1 Design Instruments . . . 99
4.2.2 Tools . . . 101
4.2.2.1 Tools for Lectures . . . 101
4.2.2.2 Tools for Labs . . . 104
4.2.3 Datasets . . . 105
4.2.3.1 Dataset and Annotation Criteria . . . 105
4.2.3.2 Using the Dataset for Pedagogical Purposes . . . 105
4.2.4 Participants . . . 106
4.2.5 Physical Location . . . 107

4.3 Phase 4 - Syllabus Implementation Methodology . . . 108
4.3.1 Theory-based classes - Lectures . . . 109
4.3.1.1 Role of the Teacher and the Students . . . 111
4.3.1.2 Duration of the Classes . . . 111
4.3.2 Practice-based classes - Lab sessions . . . 112
4.3.2.1 Role of the Teacher and the Students . . . 112
4.3.2.2 Duration of the Classes . . . 113
4.4 Phase 5 - Evaluation . . . 113
4.4.1 Triangulation . . . 114
4.4.2 One-Group Time Series . . . 115
4.4.2.1 Pretest Questionnaire and Procedures . . . 116
4.4.2.2 Mid-test - Classroom quiz . . . 117
4.4.2.3 Posttest Questionnaire and Procedures . . . 117
4.4.2.4 Statistical Analysis . . . 118
4.4.3 Classroom Observation . . . 120
4.4.4 Thematic Analysis . . . 122

5 Methods - Course Creation - "Development" 125
5.1 Phase 3 - Development . . . 125
5.2 Unit 1 - Using Metrics Critically Pt.1 . . . 127
5.2.1 Content of Unit 1 . . . 127
5.2.2 Teaching Methodology of Unit 1 . . . 130
5.2.2.1 Lesson 1 - Lecture . . . 131
5.2.2.2 Lesson 2 - Lab . . . 132
5.2.3 Assessment of Unit 1 . . . 140
5.3 Unit 2 - Using Metrics Critically Pt.2 . . . 143
5.3.1 Content of Unit 2 . . . 143
5.3.2 Teaching Methodology of Unit 2 . . . 145
5.3.2.1 Lesson 3 - Lecture . . . 145
5.3.2.2 Lesson 4 - Lab . . . 149
5.3.3 Assessment of Unit 2 . . . 156
5.4 Unit 3 - Evaluation Design . . . 158
5.4.1 Content of Unit 3 . . . 158
5.4.2 Teaching Methodology of Unit 3 . . . 160
5.4.2.1 Lesson 5 - Lecture . . . 161
5.4.2.2 Lesson 6 - Lab . . . 165
5.4.3 Assessment of Unit 3 . . . 167
5.5 Unit 4 - Ethics in Evaluation . . . 167
5.5.1 Content of Unit 4 . . . 167
5.5.2 Teaching Methodology of Unit 4 . . . 171
5.5.2.1 Lesson 7 - Lecture . . . 171
5.5.2.2 Lesson 8 - Lab . . . 172
5.5.3 Assessment of Unit 4 . . . 174

6 Results & Analysis - Course Delivery - "Implementation" 176
6.1 Phase 4 - Syllabus Implementation . . . 176
6.2 Contextual Factors in Practice . . . 176
6.3 Qualitative Results - Practice vs Lesson Plans . . . 178
6.3.1 Lesson 1 - Lecture . . . 178

6.3.2 Lesson 2 - Lab . . . 179
6.3.3 Lesson 3 - Lecture . . . 181
6.3.4 Lesson 4 - Lab . . . 182
6.3.5 Lesson 5 - Lecture . . . 186
6.3.6 Lesson 6 - Lab . . . 193
6.3.7 Lesson 7 - Lecture . . . 203
6.3.8 Lesson 8 - Lab . . . 203
6.4 Quantitative Results - Measuring Effectiveness . . . 208
6.4.1 Pretest Questionnaire Results . . . 208
6.4.2 Mid-test - Classroom quiz Results . . . 210
6.4.3 Posttest Questionnaire Results . . . 211
6.4.4 Inferential Analysis: Bayesian Linear Mixed Modelling . . . 213

7 Discussion - Integrating Classroom Data into Praxis - "Evaluation" 216
7.1 Themes and Nvivo Coding . . . 217
7.1.1 Unit 1 . . . 218
7.1.1.1 CSL - Subthemes . . . 218
7.1.1.2 RTL - Subthemes . . . 220
7.1.1.3 RE - Subthemes . . . 224
7.1.1.4 TPM - Subthemes . . . 224
7.1.2 Unit 2 . . . 225
7.1.2.1 RTL - Subthemes . . . 225
7.1.2.2 RE - Subthemes . . . 227
7.1.2.3 TPM - Subthemes . . . 228
7.1.2.4 CSL - Subthemes . . . 229
7.1.3 Unit 3 . . . 229
7.1.3.1 RTL - Subthemes . . . 229
7.1.4 Unit 4 . . . 232
7.1.4.1 RTL - Subthemes . . . 233
7.1.4.2 TPM - Subthemes . . . 236
7.1.4.3 CSL - Subthemes . . . 237
7.2 What can be taught to NLP students to encourage them to carry out HE? . . . 239
7.2.1 Unit 1 . . . 239
7.2.2 Unit 2 . . . 240
7.2.3 Unit 3 . . . 241
7.2.4 Unit 4 . . . 242
7.3 What can be drawn from TS when teaching MTQA for NLP post-graduate audiences? . . . 244
7.3.1 Unit 1 . . . 244
7.3.2 Unit 2 . . . 245
7.3.3 Unit 3 . . . 246
7.3.4 Unit 4 . . . 247
7.4 How effective is the content proposed for HE in an MT evaluation module? . . . 247
7.4.1 Learning Outcome 1 . . . 248
7.4.2 Learning Outcome 2 . . . 249
7.4.3 Learning Outcome 3 . . . 251

7.4.4 Learning Outcome 4 . . . 252
7.4.5 Learning Outcome 5 . . . 253

8 Conclusion 255
8.1 Research Questions Revisited . . . 255
8.1.1 RQ1 - What can be taught to NLP students to encourage them to carry out HE? . . . 255
8.1.2 RQ2 - What content related to MTQA is currently taught in post-graduate NLP and TS programmes, if any? . . . 256
8.1.2.1 RQ2a - What research has been published on MTQA teaching for TS and NLP? . . . 257
8.1.2.2 RQ2b - What has been taught in MTQA for NLP postgraduate students? . . . 258
8.1.3 RQ3 - What can be drawn from TS when teaching MTQA for NLP postgraduate audiences? . . . 258
8.1.4 RQ4 - How effective is the content proposed for HE in an MT evaluation module? . . . 259
8.2 Contributions . . . 259
8.3 Limitations . . . 262
8.4 Future Work . . . 264

Bibliography 266

A Systematic Literature Review Protocol 303

B Systematic Literature Review Selection 315

C Phase 1b - Curriculum Analysis - Programmes Analysed 342

D Phase 1c - Questions for Survey 347

E Phase 3 - Syllabus Development 351

F Phase 3 - Course Development 354
F.1 Lesson Plan 1 - Lecture . . . 354
F.2 Lesson Plan 2 - Lab . . . 361
F.3 Lesson Plan 3 - Lecture . . . 368
F.4 Lesson Plan 4 - Lab . . . 374
F.5 Lesson Plan 5 - Lecture . . . 385
F.5.1 Mid-test - Classroom quiz . . . 392
F.6 Lesson Plan 6 - Lab . . . 395
F.7 Lesson Plan 7 - Lecture . . . 403
F.8 Lesson Plan 8 - Lab . . . 411

G Phase 5 - Pretest Questionnaire 415

H Phase 5 - Posttest Questionnaire 419

I Inferential Analysis: Bayesian Linear Mixed Modelling **422**
I.0.0.1 Confidence in Human Evaluation 422
I.0.0.2 Confidence in Automatic Evaluation Metrics 426
I.0.0.3 Summary of results 428

J Questionnaire responses **429**

K Thematic Analysis Occurrences per Unit **435**

L Faculty Research Ethics Committee Approval **439**

List of Figures

1.1 Mixed-methods phases plan integrated with the ADDIE framework for the creation of a NLP course on MTQA 8

2.1 Illustration representing how interpretation attempts to penetrate through constraints represented as solid walls. Each solid walls represents layers that stabilise interpretations as it passes through each wall, in order to reach the MT output. 30

2.2 Theoretical rationale informed by post-positivism, aligning translation hermeneutics and critical pedagogy in dialogue, with instructional design as the structuring methodology for teaching practice. . . 36

4.1 Figure showing the order of the methodology followed for each of the remaining phases 78

4.2 Flowchart showing the five blocks, Ethical Statement (Block 0), Part 1 - Teacher Profile (Block 1), Part 2 - Human Evaluation Predictions (Block 2), MTQA Pedagogical Questions (Block 3), and Human Evaluation Teaching Recommendations (Block 4) 82

4.3 Graph showing the countries of the participants 83

4.4 Graph showing the field of the participants 84

4.5 Graph showing the activities other than teaching between participants, including research, administrative duties and other activities . 85

4.6 Distribution of educators' reported influence over the curriculum, grouped by academic background: CL/NLP and TS. The levels of influence are ordered from lowest to highest: No Influence, Minimum Influence, Some Influence, Significant Influence, and Comprehensive Influence. 86

4.7 Bar chart showing the number of educators teaching at different academic levels, grouped by CL/NLP and TS. Each bar represents the frequency of responses for a given level, including Undergraduate, Postgraduate, Doctoral, Continuous Professional Development, Online Courses, and Other. Respondents could select multiple levels. . . 87

4.8 Teaching modalities reported by educators in CL/NLP and TS. The chart shows the number of respondents teaching in each format: In-person, Hybrid (a mix of online and in-person), and Online 88

4.9 Number of respondents who have taught MT quality assessment, grouped by disciplinary background (CL/NLP and TS). 89

4.10 Perceived importance of teaching MT quality assessment, grouped by disciplinary background (CL/NLP and Translation Studies). Responses were collected on a 5-point Likert scale: 1 – Not at all important, 2 – Slightly important, 3 – Moderately important, 4 – Very important, and 5 – Extremely important. 90
4.11 Trends foreseen in HE for MT, grouped by disciplinary background (CL/NLP and TS). Categories include customised evaluation for specific domains, user experience, context, ethics, multimodal evaluation, crowdsourced evaluation, and other. Multiple selections were allowed. 91
4.12 Types of MT evaluation methods taught, grouped by disciplinary background (CL/NLP and TS). Categories include Automatic, Human, and Semi-Automatic evaluation. Multiple selections were allowed. 92
4.13 Perceived importance of teaching students about MT evaluation design, grouped by disciplinary background (CL/NLP and TS). Responses are shown on a 5-point Likert scale ranging from "Not at all important" to "Extremely important". 93
4.14 Importance of including evaluation metrics in the academic curriculum, grouped by AEMs and HE and disciplinary background (CL/NLP and TS). Responses are shown on a 5-point Likert scale ranging from not at all important to extremely important 94
4.15 Pedagogical aspects perceived as lacking in the teaching of MT quality assessment, grouped by field (CL/NLP and TS). Respondents selected from: "Allocated Hours", "Curricular Structure", "Faculty Expertise" and "Development", "Scalability of Teaching Methods", and "other". Multiple selections were permitted. 95
4.16 Suggested content for a Human Evaluation module in MT quality assessment, grouped by field (CL/NLP and Translation Studies). Responses reflect a range of potential content areas, with multiple selections being allowed. 96
4.17 Preferred formats for training in MT quality assessment, grouped by field (CL/NLP and Translation Studies). Respondents selected from options such as workshops, short courses, academic modules, and online formats. Percentages reflect the proportion of selections within each group. 97
4.18 An illustration of how the course blueprint elements from Phase 2 - Design result in a syllabus and a course for Phase 3 - Development . 98
4.19 Definitions and directions of each element of an IM that the ID professional must follow in the design phase. 100
4.20 Requirements for showing how the needs analysis results in the sequencing of content and instruction structure 101
4.21 Rationale explaining the core principle of lesson structures and sequencing 101
4.22 Heading of the Lesson Plan detailing basic descriptors of the lesson before the instructions 102
4.23 Heading of a Lesson Plan detailing basic descriptors of the lesson before the instructions 102

4.24 Slide illustrating a conversation between a blue-haired character and a mentor-like character talking about the need for measuring the quality of MT systems . . . 103
4.25 Slide showing a scenario of evaluation with different characters, including a woman character as a team lead, a scientist with a disability and a non-binary scientist. . . . 103
4.26 Interface of Vevox, where the teacher can create multiple choice questions, scale-based questions and word clouds. . . . 104
4.27 Differences between traditional and dialogic lectures, with the traditional showing the teacher in a vertical relationship with students, while the dialogic shows a horizontal relationship, where knowledge is built together . . . 109
4.28 Figure illustrating the integration of a One-Group Time Series Quasi-Experiment and Classroom Observations to examine training effectiveness and students' confidence progression in MTQA, where qualitative data validates and complements quantitative findings to answer RQ1 and RQ4 . . . 114
4.29 Figure illustrating the phases of the One-Group Time Series Quasi-Experiment showing the pretest, midtest and posttest converging the measurements into statistical results . . . 115
4.30 Variables of the study. The independent variable is the training intervention, operationalised as three timepoints (T1–T3). At each point, confidence in HE and AEMs was measured. . . . 118

5.1 Simplified Syllabus containing the learning outcomes, units, learning activities and assessment. . . . 126
5.2 Simplified description of topics covered, learning activities and assessment of Unit 1. . . . 128
5.3 Simplified description of topics covered over Lesson 1, delivered as a lecture. . . . 130
5.4 Simplified description of tasks covered over Lesson 2, delivered as a lab session. . . . 132
5.5 Interface of Task 1, showing segments of the ST and the TT . . . 133
5.6 Interface of Task 1, showing the categories of adequacy, fluency and error classification with questions to inform the quality assurance and drop-down menus . . . 134
5.7 Part of the interface of Task 2, showing columns "T1" (translator 1) and "T2" (translator 2) to represent scores drawn from Task 1. . . . 135
5.8 Interface of Task 2, a matrix illustrating the agreement between two raters (T1 and T2) across four categories, including observed agreement, chance agreement, and a Kappa value. The bottom rows show category-specific totals and proportions. . . . 135
5.9 Interface of the teacher's version of the document sheet of Task 1, with colours identifying the MT systems. . . . 137
5.10 Interface of Task 3, where users assess the comprehensibility of MT output by categorising multilingual reviews based on their sentiment or descriptive nature, where evaluators can choose from positive, negative or neutral judgements from a drop-down menu. . . . 137

5.11 Interface of Task 3 for the teacher only, reference columns for instructors to guide students' questions during the comprehensibility evaluation task. It provides the correct answers and corresponding reasons, enabling teachers to clarify doubts and support students in understanding the quality and clarity of MT. 138
5.12 Interface of Task 4. It shows a Likert scale questionnaire where users rate their experience with the MT output from the comprehension task. Questions assess clarity, satisfaction, overall comprehensibility, potential for future understanding, recommendation likelihood, and the overall usability of the translated reviews. 139
5.13 Interface of Task 5. Students evaluate and select their preferred translation or indicate if both are equally good. The task focuses on assessing MT outputs from an end user perspective, with no specific criteria given. 139
5.14 Simplified description of topics covered, learning activities and assessment of Unit 2. 143
5.15 Simplified description of topics covered over Lesson 3, delivered as a lecture. 146
5.16 Interface of Qualtrics with explanations related to Task 1 149
5.17 Interface of Qualtrics for the evaluation of adequacy and fluency . . . 149
5.18 Question on Qualtrics about confidence when evaluating the sentence 150
5.19 Instructions on how to evaluate a MT sentence highlighted in yellow used the context span approach 151
5.20 Question on Qualtrics describing the task of a document-level evaluation activity 152
5.21 Example of a numbered sentence highlighted in yellow for students to evaluate 152
5.22 DA scale presented to students for rating adequacy and fluency . . . 153
5.23 DA scale to rate the full document in respect of adequacy. 154
5.24 DA scale showing four increments (0, 25, 50, 75, 100), two labels ("None" and 'All of It"), no custom position with scaling showing the selection upon clicking and a value shown to evaluators (in this case, "33"). 154
5.25 Simplified description of topics covered, learning activities and assessment of Unit 3. 159
5.26 Simplified description of topics covered over Lesson 5, delivered as a lecture. 161
5.27 Simplified description of topics covered, learning activities and assessment of Unit 4. 168
5.28 Pedagogical model of translation process automation, illustrating eight levels of increasing automation and associated translation types, from fully manual human translation to raw machine translation. 169

6.1 Word cloud showing the frequency of terms used in 15 participant responses about the strengths of document-level evaluation, with larger words indicating higher frequency 190

6.2 Word cloud showing the frequency of terms used in 14 participant responses about the challenges of document-level evaluation, with larger words indicating higher frequency . . . 191
6.3 Example of the first page of the evaluation sheet, which contained a checklist and note taking spaces for each task . . . 194
6.4 Example of part of a brief, which contained a scenario, type of MT system requested for development/evaluation and the context of the current system each group had to adapt . . . 195
6.5 Example of the first page of the instructions, which provides explanations on what factors to consider for each task in Lesson 6 . . . 196
6.6 Example of one of the case studies crafted for Task 1, containing preposterous claims, inaccurate reporting or questionable evaluation practices . . . 204
6.7 Mean *HEConfidence* across time, by participant background (NLP vs TT). . . . 214

7.1 Figure detailing Phase 5 - Evaluation. It considers the process of coding themes from classroom data. The classroom observation was performed across four units, to collect data from reflective reports, evaluation plans, and case studies. The data informs the thematic categories, resulting into themes from the SRL, which address RQ1 and RQ4. . . . 217
7.2 Heatmap illustrating the distribution of subthemes identified from TPM, RE, CSL, and RTL across four units. The intensity of the shading reflects the frequency of each theme's occurrence, with the highest observed frequency being 35 instances for a single theme. . . . 219
7.3 Heat map illustrating the distribution of subthemes identified from TPM, RE, CSL, and RTL across Unit 1. The intensity of the shading reflects the frequency of each theme's occurrence, with the highest observed frequency being 17 instances for a single theme. . . . 220
7.4 Heat map illustrating the distribution of subthemes identified from TPM, RE, CSL, and RTL across Unit 2. The intensity of the shading reflects the frequency of each theme's occurrence, with the highest observed frequency being 19 instances for a single subtheme. . . . 225
7.5 Heatmap illustrating the distribution of subthemes identified from TPM, RE, CSL, and RTL across Unit 3. The intensity of the shading reflects the frequency of each theme's occurrence, with the highest observed frequency being 35 instances for a single subtheme. . . . 230
7.6 Heatmap illustrating the distribution of subthemes identified from TPM, RE, CSL, and RTL across Unit 4. The intensity of the shading reflects the frequency of each theme's occurrence, with the highest observed frequency being 18 instances for a single subtheme. . . . 232

I.1 Mean *HEConfidence* across time, by participant background (NLP vs TT). . . . 423
I.2 Mean *HEConfidence* across time, by participant background (NLP vs TT). . . . 424
I.3 Mean AEMConfidence across time, by participant background (NLP vs TT). . . . 427

List of Tables

3.1 Data Synthesis, Extraction, Summarisation, and Publication Strategy 42
3.2 Number of studies for each Stage of the SRL, including the total number, followed by the removal of duplicates (first round), and the application of inclusion and exclusion criteria (second round). 43
3.3 Description of SRL-Questions matched with themes and justification for the topics that must be discussed in each theme 46
3.4 Total of studies attributed to themes in Stage 1 46
3.5 Studies classified by thematic focus. An "x" marks the presence of a theme within each study: Rationale (RTL), Resources (RE), Modality (MOD), Teaching Procedures and Methodologies (TPM), and Curriculum and Syllabus (CSL). Lines 1–5 indicate single-theme studies, while lines 6–12 represent multithemed studies addressing two or more categories. 48
3.6 Pedagogical contributions of studies under Rationale organised by authors and respective reference numbers 50
3.7 Pedagogical contributions of studies under Resources organised by authors and respective reference numbers 54
3.8 Pedagogical contributions of studies under Modality organised by authors and respective reference numbers 57
3.9 Pedagogical contributions of studies under Resources organised by authors and respective reference numbers 59
3.10 Pedagogical contributions of studies under Curriculum and Syllabus organised by authors and respective reference numbers 61
3.11 Pedagogical contributions of studies under Curriculum and Syllabus organised by authors and respective reference numbers 64

4.1 Descriptive statistics for hours spent on teaching by field 84
4.2 Descriptive statistics for *years* of experience teaching MT quality assessment, grouped by field 92

6.1 Student mapping of Triple Bottom Line (TBL) elements in two sustainability papers, with notes on their relevance to MTQA. 206
6.2 Descriptive statistics for HE and AEM confidence at Pre-test 208
6.3 Frequency distribution of HE confidence scores at Pre-test 209
6.4 Frequency distribution of AEM confidence scores at Pre-test 209
6.5 Descriptive statistics for HE and AEM confidence at Mid-test 210
6.6 Frequency distribution of HE confidence scores at Mid-test 210
6.7 Frequency distribution of AEM confidence scores at Mid-test 211
6.8 Descriptive statistics for HE and AEM confidence at Post-test 211

6.9 Frequency distribution of HE confidence scores at Post-test 212
6.10 Frequency distribution of AEM confidence scores at Post-test 212
6.11 Summary of fixed effects from the linear mixed model for HEConfidence. The table reports each predictor, the numerator and denominator degrees of freedom used in the F-test, the F-value, and the associated significance level (p-value). 213
6.12 Estimated marginal means of HEConfidence by Time 215

7.1 Categorisation of themes and their corresponding subthemes identified in the analysis. 218
7.2 Occurrences and percentages of subthemes coded in Unit 1. 221
7.3 Occurrences and percentages of subthemes coded in Unit 2. 226
7.4 Occurrences and percentages of subthemes coded in Unit 3. 229
7.5 Occurrences and percentages of subthemes coded in Unit 4. 233
7.6 NLP Participants reflections from Reflective Report #1 present in Lesson 2, Unit 1 where they demonstrated willingness to use HE or positive attitudes towards HE. 240
7.7 NLP Participants reflections from Reflective Report #2 present in Lesson 4, Unit 2 where they demonstrated willingness to use HE or positive attitudes towards HE. 241
7.8 NLP Participants reflections from the evaluation planning assignment present in Lesson 6, Unit 3 where they demonstrated willingness to use HE or positive attitudes towards HE 241
7.9 NLP Participants reflections from the post test questionnaire filled at the end of Lesson 8, Unit 4 where they demonstrated willingness to use HE or positive attitudes towards HE 243
7.10 NLP participants' reflections from Unit 1 on how translation studies resources (e.g., evaluation criteria and judgement elicitation practices) influenced their understanding of NLP evaluation 244
7.11 NLP participants' reflections from Unit 2 on hands-on engagement with AEMs, scaling practices, and the effect of context in translation evaluation 246
7.12 NLP participants' reflections from Unit 4 on ethical considerations, transparency, reproducibility, and responsible evaluation frameworks inspired by translation studies 247
7.13 Participants' views on HE application, interaction with AEMs, and evaluation frameworks within MT development (LO1), based on themes from Section 7.1 extracted from reflective reports, assignments and questionnaires. 248
7.14 Participants' views on AEM application and weaknesses within MT development (LO2), extracted from reflective reports, assignments and questionnaires. 250
7.15 Participants' views on evaluation frameworks, UX/UI, human-centred perspectives, and ethical considerations in MT development (LO3), extracted from reflective reports, assignments and questionnaires . . . 251
7.16 Participants' reflections on reproducibility, domain use, and ethical evaluation design in MT development (LO4), extracted from reflective reports, assignments and questionnaires. 253

7.17 Participants' views on reliability and the consequences of poor MTQA within MT evaluation reporting (LO5), extracted from reflective reports, assignments and questionnaires. 254

C.1 Universities and Master's programmes analysed 342

I.1 Summary of fixed effects from the linear mixed model for HEConfidence. The table reports each predictor, the numerator and denominator degrees of freedom used in the F-test, the F-value, and the associated significance level (p-value). 423
I.2 Estimated marginal means of HEConfidence by Time 425
I.3 Estimated marginal means of HEConfidence by Time and Background 425
I.4 Linear mixed model results for AEMConfidence 426
I.5 Estimated marginal means of AEMConfidence by Time 426
I.6 Estimated marginal means of AEMConfidence by Time and Background 427

K.1 Occurrences and percentages of subthemes coded in Unit 1. 435
K.2 Occurrences and percentages of subthemes coded in Unit 2. 436
K.3 Occurrences and percentages of subthemes coded in Unit 3. 437
K.4 Occurrences and percentages of subthemes coded in Unit 4. 438

List of Abbreviations

Abbreviation	Meaning
ADDIE	Analysis, Design, Development, Implementation, Evaluation
AEM	Automatic Evaluation Metric
CSL	Curriculum and Syllabus
DQF	Dynamic Quality Framework
EMT	European Master's in Translation
HCI	Human-Computer Interaction
HE	Human Evaluation
IAA	Inter-Annotator Agreement
ID	Instructional Design
LLM	Large Language Model
LMS	Learning Management System
MT	Machine Translation
MOD	Modality
MTQA	Machine Translation Quality Assessment
NLP	Natural Language Processing
QA	Quality Assessment
RE	Resources
RTL	Rationale
SME	Subject Matter Expert
SRL	Systematic Review of Literature
SRL-Q	Systematc Review of Literature Research Questions
ST	Source Text
TQA	Translation Quality Assessment
TH	Translation Hermeneutics
TPM	Teaching Procedures and Metrics
TS	Translation Studies
TT	Translation Technology
Tt	Target Text
UI	User Interface
UX	User Experience

Structured by ADDIE, Guided by Freire: Developing Critical MT Quality Assessment Training

João Lucas Cavalheiro Camargo

Abstract

This research develops and evaluates a pedagogical intervention for teaching Machine Translation Quality Assessment (MTQA) to postgraduate students in Natural Language Processing (NLP). While Translation Studies (TS) has a long tradition of exploring MTQA, MTQA pedagogy is underdeveloped in NLP education. In both academic and industry contexts, MTQA is most commonly carried out using Automatic Evaluation Metrics (AEMs). These metrics are valued for their speed, low cost, and scalability, but they fail to capture important linguistic, contextual, and ethical aspects of translation quality. Human Evaluation (HE), while capable of providing more reliable and nuanced insights, is underused because it is complex, resource-intensive, and time-consuming. As Neural Machine Translation (NMT) continues to produce mistranslations, omissions, and biases, and Large Language Models (LLMs) raise new risks such as hallucinations and instability across system updates, the need for rigorous evaluation practices becomes even more critical.

To address this imbalance, this study draws on the ADDIE instructional design framework and Freirean critical pedagogy to structure a training course on MTQA. A systematic review of literature and a survey of educators identified gaps in how MTQA is taught and reported. Based on these findings, a syllabus and course were designed, combining theoretical foundations with hands-on practice in HE, AEMs, evaluation design, and ethics. The course was implemented with postgraduate NLP students, who participated in lectures, lab-based evaluation tasks, and critical reflection on the methodological and ethical dimensions of MTQA.

Effectiveness was assessed through a mixed-methods design, including pre-, mid- and post-tests analysed with Bayesian linear mixed modelling, classroom observation, and thematic analysis of student reflections. The results show that the students increased their confidence in conducting HE, developed a greater critical awareness of the the strengths and weaknesses of each type of evaluation, and demonstrated the ability to apply rigorous evaluation practices to MT systems.

By designing, delivering, and evaluating this course, the study demonstrates that critical MTQA training is both feasible and effective. It contributes a structured pedagogical model for embedding responsible and transparent evaluation practices into NLP curricula, promoting more human-centred and accountable approaches to the development of systems focused on AI-driven translation.

Chapter 1

Introduction

1.1 Context

There is a significant growth in content that needs to be translated (Moorkens, Castilho, Gaspari and Doherty, 2018), and there are millions of users of machine translation (MT) systems for different tasks (Way, 2018). It shows that MT is in demand, and that it is used to complete translation tasks according to the needs of these users. However, MT systems are far from perfect, and MT is a Natural Language Processing (NLP) task that is far from solved (Moorkens et al., 2024b). There is continuous effort from companies and academic researchers to verify and improve the quality of MT systems (Way, 2018). Designed to measure how well or poorly they perform, the quality assurance (QA) process for MT can be called MT quality assessment (MTQA).

In Translation Studies (TS), researchers have examined TQA extensively over the years. According to House (2014), measuring the quality of translation requires a careful analysis involving language and expertise in translation. When discussing the state of TQA in TS education that is increasingly technologised and focused on MT, Doherty et al. (2018) explain how important it is to integrate TQA into TS academic curricula using state-of-the-art evaluation metrics, as graduates face challenges that require skills related to MTQA in the industry, usually those that combine knowledge of state-of-the-art metrics with methodological rigour in evalua-

tion processes. Combining both, it would be possible to achieve smoother workflows and encourage diverse approaches in MTQA (Doherty et al., 2018).

Over time, TS has evolved and formalised how evaluation occurs in MT systems alongside NLP, as can be seen in Castilho et al. (2018). Castilho et al. (2018) explain that assessing the quality of the output of MT systems can be performed by humans or by machines. The former, human evaluation (HE), is ideal, but can be slow and expensive, while the latter employs automatic evaluation metrics (AEM), which are cheap and quick. This difference leads to the adoption of AEMs over HE, in the process of MTQA. However, AEMs, while convenient, have flaws. Although AEMs provide useful quantitative information about systems, they cannot provide a detailed analysis of linguistic phenomena, something only HE can do (Castilho and Caseli, 2024). Way (2018) recommends combining AEMs with HE, as overreliance on AEMs introduces weaknesses in the process of diagnosing systems and generalising results in evaluation campaigns. Moorkens (2022) explains that in campaigns, called shared tasks, some results are reported so as to imply that MT can substitute human translators if the MT output matches a reference translation according to these AEMs. Because of that, Doherty (2017) suggests that solid methodological control then becomes essential to mitigate these issues.

The two contemporary paradigms of MT have their own weaknesses, amplifying the problems reported previously. The first paradigm is Neural MT (NMT). NMT systems are built using neural networks (Koehn, 2020, p. 33). These systems can produce grammatical errors, mistranslated words or constructions (Koehn and Knowles, 2017), and incomplete output. They can mistranslate named entities, include errors specific to part of speech and word order (Guerreiro et al., 2023), which can be attributed to the phenomenon named hallucination. This type of taxonomy is attributed to this type of phenomena that occurs in different ways in NMT (Raunak et al., 2021). They can produce gender bias due to word embeddings connecting professions with specific genders (Prates et al., 2020). Relatedly, they can also introduce biases with different pronouns (Lauscher et al., 2023), demonstrat-

ing how MT handles pronoun translation poorly, harming non-binary individuals and individuals of other gender identities. Furthermore, these errors are dangerous when they occur in sensitive domains, such as medical and legal domains (Vieira et al., 2020). The second paradigm involves Large Language Models (LLMs). While they are not MT systems, they have the same basis in the transformer architecture used to translate (Brown et al., 2020), and this rationale for LLM translation will be adopted in this thesis. LLMs can also produce hallucinations (Bowman, 2024), which can be understood as generated content/output that does not correspond to the source content/input provided (Huang et al., 2025). Most importantly, since LLMs are a relatively recent technology, existing methods and paradigms in NLP are being tested to understand their effectiveness (Bowman, 2024). An added difficulty is that systems are constantly changing and updated the moment experiments are first conducted (Briva-Iglesias et al., 2024).

Both NMT and LLMs are built using Machine Learning, which Kühl et al. (2020) define as a set of methods for learning patterns in existing data that generate analytical models embedded within larger information technology systems. Machine learning, however, sometimes is used interchangeably with the concept of Artificial Intelligence (AI). AI applications either emphasise human-like decisions or act as an intelligent agent to achieve desired outcomes (Kühl et al., 2020). In the context of this thesis, these definitions will be used interchangeably unless otherwise specified, whether it refers to machine learning techniques or the ability of a machine to make a human-like decision. In the face of the previously mentioned risks and the difficulty of asserting stable and rigorous analyses of these systems, working with AI-based systems requires responsibility. Hence, when using, developing or evaluating AI-based systems, developers must "recognise the need for design and engineering approaches that ensure the safe, beneficial, and fair use of AI technologies" (Dignum, 2020, p. 216). In this manner, the measurement of quality must not rely entirely on the machines if the purpose of use of the system is to produce content that is long lasting or of critical importance. Marie et al. (2021) conducted a meta-evaluation

of 769 papers published between 2010 and 2020 in the area of MTQA, and demonstrated a reliance of studies on the metric BLEU (Papineni et al., 2002), in addition to low use of HE alongside BLEU and other AEMs, indicating a problem in the measurement of the performance of systems and the reporting of results in MTQA research. Marie et al. (2021) evidenced the issue of overreliance on AEMs and lack of rigorous and reliable evaluation methodology highlighted in MTQA literature e.g. (Doherty, 2017; Way, 2018). Kocmi et al. (2023) recommends replacing traditional n-gram AEMs such as BLEU with pre-trained neural AEMs such as COMET, which correlate more strongly with human judgements. However, these models also have limitations, for example, Amrhein and Sennrich (2022) show that COMET is insufficiently sensitive to errors with numbers and named entities.

Läubli et al. (2020) emphasise that HE is essential to affirm how valid results are, affecting what claims can be made from experiments. The authors recommend a set of guidelines to support a more transparent and comprehensive HE which include: choosing professional translators as raters; evaluating documents and not sentences; always evaluating the target language alongside the source language; not heavily editing reference translations to make them more fluent; using original source texts. Other recommendations include ensuring high inter-annotator agreement (IAA) and disclosing the number of assessments (Graham et al., 2019); evaluating multiple human translations; an MT system must be compared to all translation provided by humans in the evaluation and; the rater should match the end-user of the MT system being evaluated (Gilbert, 2023). These guidelines are examples of thorough MT evaluation that can be applied and reported in order to achieve more rigorous MTQA, as recommended by Marie et al. (2021).

As AI technologies greatly impact humans and are increasingly present in society, it is important not to assign important tasks such as evaluation entirely to machines (Floridi et al., 2018). Because of the challenge and progress made in MTQA, "we need to be aware of the impact of the choices we make when we design, implement, or use technology" (Moorkens, 2022, p. 122). For example in MT, machines

lack contextual and experiential world knowledge and can learn from (risky) human biases (Reijers and Dupont, 2023; Castilho and Knowles, 2024). Furthermore, while regulation can mitigate some of the dangers, it does not always match the speed of technological progress. Currently, to mitigate ethical issues and inadequate measurement of the performance of AI technologies, engineers and developers can adopt ethical behaviour when developing and diagnosing systems (Moorkens, 2022, p. 122). This accentuates the complexity of MTQA, and as a result, asserts designing accountable, responsible and transparent HE suitable to a doctoral investigation.

1.2 Motivation

NLP is a field that involves translation in its dedicated tasks. As AI-driven translation presents the risks mentioned previously, designing and developing systems must include ethical behaviours that mitigate biased and risky content reaching end users. The literature provides the rationale that overreliance on one type of metric leads to poor evaluation design (Marie et al., 2021) and lack of training leads to workflows that do not fully use the appropriate or state of the art techniques, tools and metrics (Doherty, 2017). While the literature advocates that MT end users become more critical and responsible (Moorkens et al., 2025), I contend that developers can also be more critical and responsible. Considering that developers are relevant stakeholders and sometimes users of these systems, it is relevant to ensure integrity in the research and use of AI-driven translation systems through training (Dignum, 2020, p. 217). NLP graduates who receive critical MTQA training can assert themselves as critical users who decide to responsibly pursue the path of working with automated translation. **In this context, my motivation** is training NLP students about critical MTQA, using resources and scholarship from the field of TS and Translation Technology (TT).

In the context of NLP education, the field has hosted workshops starting in 2002 and having different editions in 2005, 2008, 2013 and 2021, accumulating a total of 83 papers published over the years (ACL Anthology, 2023). Educational research in

the field was conducted as early as 2002 with the Natural Language Toolkit, the first open-source toolkit designed for teaching NLP in Python; with general teaching of NLP tasks, including MT, in its curriculum (Brew et al., 2005; Hearst, 2005; Klein, 2005; Uszkoreit et al., 2005). The deficit in more resources and pedagogical studies represent the need for more scholarship from MTQA pedagogy in NLP education.

1.3 Research Questions

Considering the context presented and the motivation described, the following research questions (RQ) have been formulated:

RQ1. - What can be taught to NLP students to encourage them to carry out HE?

If TS already has a tradition of teaching MTQA, it is appropriate for NLP students who will carry out the evaluation of MT systems to consider methods from TS. In that manner, if NLP students aim to be MT developers, being trained to devise an evaluation plan is valuable as they can learn to perform a strong HE of MT through the appropriate choice of procedures. Analysing, designing, developing, implementing and evaluating this training would enable NLP students to ensure they evaluate MT output using HE while adhering to principles of responsibility and transparency, to mitigate risks such as biased or harmful output.

RQ2. - What content related to MTQA is currently taught in post-graduate NLP and TS programmes, if any?

This question can be split into two sub questions:

RQ2a - What research has been published on MTQA teaching for TS and NLP?

To answer this RQ, I devised a Systematic Review of Literature (SRL) in order to identify the current pedagogical scholarship on MTQA, both for TS and NLP. The SRL has two stages to address the two fields, first TS and then NLP. As previously stated, it is known that TS has a strong commitment to teaching the use of MT and its assessment (Rico and González Pastor, 2022), so it is investigated to map

and identify its practices. For NLP, the aim is to identify existing MTQA teaching studies or if MTQA teaching has been reported at all in the field.

RQ2b - What has been taught in MTQA for NLP postgraduate students?

As no major studies were found in RQ2a, it could mean that MTQA has been taught in programmes but its teaching may not have been reported academically. Therefore, to answer RQ2b, I performed a curricular analysis of Master's programmes in NLP in Europe, to complement RQ2a's findings.

RQ3 - What can be drawn from TS when teaching MTQA for NLP postgraduate audiences?

Given that there is little reporting of a teaching tradition of MTQA for NLP Master's students (to be addressed in RQ2a and RQ2b), to partially answer question RQ3, I devised a survey. The survey inquires what MTQA teachers in the fields of TS and NLP believe to be necessary for HE content on MTQA to complement the current curriculum and what they consider to be ideal factors for training. Based on the responses, I designed a syllabus containing the suggestions from this phase and the results from RQ2a and RQ2b. This led to the creation of a syllabus and a full course with defined content, learning activities and assessment. Finally, I implemented and evaluated this course by collecting student data in the process. In particular, I looked for student data that contains evidence of willingness to use TS materials and tools in their practice and assignments, so that RQ3 can be fully addressed. By TS materials, I mean what is commonly used in the field, not necessarily implying origin of creation or implementation.

RQ4 - How effective is the content proposed for HE in an MT evaluation module? I evaluate the effectiveness of the course by using Instructional Design (ID). ID processes involve assessing the implementation of courses, useful for RQ4, because it is "essential for trying to understand why an instructional system works or not, and what other factors (internal and external) are at work within the learning environment" (Seel et al., 2017, p. 72). The evaluation examines students'

attitudes throughout the learning process, such as expectations of the course. I accomplish it by adopting participative educational ethnography (Bortoni-Ricardo, 2008) and reflective teaching practice (Freeman, 1998).

To answer the research questions and to structure this thesis, I adopt a pedagogical structure using the ADDIE framework. The framework refers to the stages of **a**nalysis, **d**esign, **d**evelopment, **i**mplementation and **e**valuation (Morrison et al., 2019) in the creation of training for academic and/or corporate purposes. Thus, this doctoral research is organised in five phases, which can be seen in Figure 1.1.

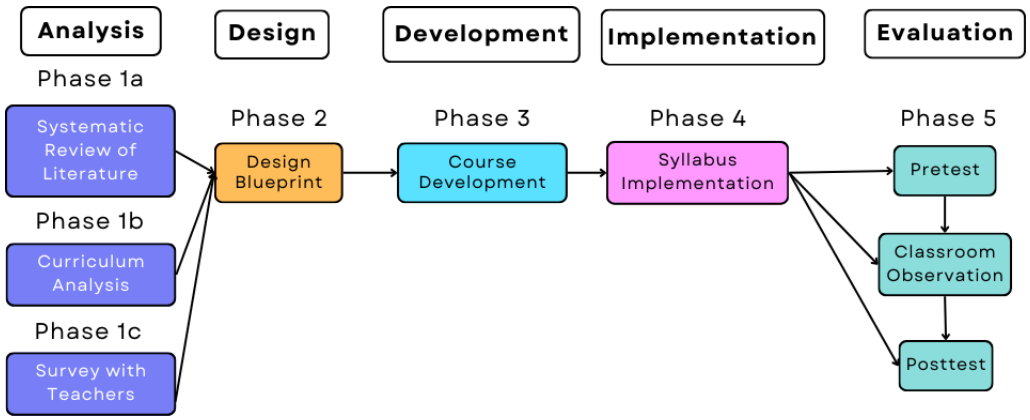


Figure 1.1: Mixed-methods phases plan integrated with the ADDIE framework for the creation of a NLP course on MTQA

1. **Analysis** - **Phase 1a** - Conduct a SRL to identify gaps (Petticrew and Roberts, 2008) in the reporting of HE in MTQA within NLP. Additionally, draw from resources for NLP (Newman and Gough, 2020) and from TS, considering its tradition of conducting and teaching translation quality assessment. This phase addresses RQ2a.

2. **Phase 1b** - Perform a qualitative curricular analysis of NLP master's programmes to assess the presence of MTQA, particularly HE. This phase addresses RQ2b, and complements the previous phase in identifying how HE of MT has been taught in NLP.

3. **Phase 1c** - Conduct a survey involving TS and NLP educators to identify crucial aspects of HE training for MTQA. It aims to identify what aspects of

HE are most important for training NLP master's students, and also aims to ascertain the most suitable training format (e.g. summer school, massive open online course (MOOC), workshops) to implement the MTQA course with the priorities given by the teachers, in order to partially address RQ3.

4. **Design - Phase 2** - Create a design blueprint by choosing the tools, teaching approaches, datasets, location, ideal participants, in order to create a syllabus and a course, leading to Phase 3.

5. **Development - Phase 3** - Create the course, establishing the methodology for my role as the teacher to adequately approach the content, teaching methodology, and assessment in order to implement the course.

6. **Implementation - Phase 4** - Implement the course created on Phase 3. This phase provides contextual information about delivery, such as the students and how they were recruited. In this stage, I, as the teacher, record the delivery of the training through research notes and student assignments. Through reflective practice, I identify if and how the practice diverged from the theory, contrasting the course planned in Phase 3 and establishing practices that converge or deviate from the literature seen in the SRL.

7. **Evaluation - Phase 5** - Analyse the training's effectiveness by examining questionnaires, notes and assignments to quantitatively and qualitatively analyse the classroom data. It is done to comprehend students' attitudes, expectations, and willingness to use techniques, tools, and resources learnt at the MTQA course. The primary objective is to validate the achievement of the learning outcomes (LO) by triangulating quantitative and qualitative data supporting the change in confidence levels and willingness to adopt the tools and approaches learnt in the course. This phase fully addresses RQ1, RQ3, and RQ4.

The thesis is divided into chapters that follow the phases of ADDIE.

1.4 Thesis Structure

The thesis is divided into chapters that combine educational philosophy with the framework of ADDIE.

Chapter 2 - Pedagogical Rationale - Provides the theoretical rationale based on translation and education. This choice is made with the assumption that teaching should combine the epistemological conceptualisation of the subject matter and the pedagogical underpinnings of the teaching. In this case, the theory of the object of study is quality assurance of translation. It provides theoretical elements that support and justify the conceptualisation and operationalisation of MTQA. The pedagogical orientation explains how the course material and design (theory) and delivery (practice) are understood. It is also anchored on a discussion of foundations of education, mainly to adopt critical pedagogy elements, based on Freire (1996).

Chapter 3 - Systematic Review of Literature - "Needs Analysis" - The purpose of this chapter is to both find a research gap related to the RQs, as discussed in previous sections and to identify the resources available in TS. Thus, this chapter provides a review of literature using an SRL in addition to a curricular analysis to complement the results of the SRL. This chapter addresses RQ2a and RQ2b.

Chapter 4 - Methods - "Design" - Provides the methodology for all phases. It contains the methods and findings of the Phase 1c survey. It also contains the methods of Phase 2, used to identify, select and organise the design blueprint that is essential for developing a course in Phase 3. It contains the methods for Phase 4, where the teaching methodology is defined for lectures and lab sessions. It also contains the quantitative and qualitative methods for Phase 5. For the quantitative component, it describes the design of the One-group Time Series experiment, to show how the participants had their learning measured and what statistical tests were chosen to analyse the data. For the qualitative component, it describes data collection through classroom observation, explaining how I position myself in the re-

search. Finally, it describes how the quantitative and qualitative data were analysed through triangulation. This chapter partially addresses RQ3.

Chapter 5 - Methods - Course Creation - "Development" - Provides the planning of a course based on the needs analysis (Phases 1a, 1b, 1c) and the design blueprints (Phase 2). It describes the structure of the course and, using the literature, justifies the selection of content and tasks for each lesson. It contains four units, their content, the teaching methodology and how students would be assessed.

Chapter 6 - Data Results & Analysis - Course Delivery - "Implementation" - Provides a description of the qualitative and quantitative data. It describes the qualitative data, that is, what occurred in practice, based on classroom observation and research notes. Through reflection, it also connects my observation as the teacher with the literature for a preliminary analysis. For the quantitative data, it describes the results of the statistical tests and draws preliminary conclusions based on them.

Chapter 7 - Discussion - Theory vs Practice - "Evaluation" - Provides a synthesis through a thematic analysis, unifying all the qualitative data collected over the course through codes. Then, it converges the results of through triangulation to fully address RQ1, RQ3, and RQ4.

Chapter 2

Epistemological Rationale

2.1 Pedagogical Rationale

The purpose of this chapter is to discuss the foundational definition of concepts that are primary to this doctoral thesis. Theory in this thesis is used as "the lens and focus" (Creswell and Clark, 2017, p. 74) of the research questions. That means the theory impacts both the object of study and how I operationalise it. In terms of discipline-specific rationale, this thesis is situated within research-based Instructional Design (ID), where "Design experiments are based on learning theories, and thus, contribute to their consolidation" (Seel et al., 2017, p. 109-110). This relationship between rationale and training design draws from specific pedagogical orientations within the development of curricula aimed at a specific topic (Zierer and Seel, 2012, p. 9). For example, this type of relationship has been adopted in TS of combining discipline-specific frameworks with pedagogical theories by the work of Hurtado Albir (2007).

Specifically, in this chapter, the discipline-specific framework defines and demonstrates how to operationalise translation and quality assessment, following scholars from TS. The pedagogical rationale, meanwhile, will guide the design of activities and will influence the implementation of the training, that is, the teaching practices. Furthermore, considering this is a mixed-methods study involving different disciplines, the use of theory provides a common understanding about the study with

different stakeholders (such as examiners) (Creswell and Clark, 2017) and provides methodological elements for describing and organising data, which are described further in the next section.

2.2 Theoretical Assumptions and Methods

Accounting for the needs of this interdisciplinary research, I drew methods from different fields, particularly TS, NLP and education. TS as a field developed new research methods and reframed methods that go beyond the traditional (Saldanha and O'Brien, 2014) to navigate complex scientific investigations. In NLP, there is an understanding that research consists of experiments[1]. Meanwhile, educational research is understood as a "complex, subtle, challenging, thoughtful activity and often a messier process than researchers would like it to be" (Cohen et al., 2018, p. 3). This alludes to the dynamic and ever-changing nature of research in education. On one hand it considers the systematic organisation, planning, validation and evaluation of research. On the other hand, it considers flexibility to adapt methods towards what is feasible and possible.

Research paradigms vary from their understanding of how knowledge can be constructed and understood. In my view, a research paradigm is the understanding of how a scientific community plan, effect and report research (Kuhn, 1962). In this research, because I navigate various communities, I believe I must adopt interdisciplinarity. It considers the assumptions and methods from related paradigms and communities. This understanding leads me to carefully navigate and apply **mixed-methods**, to incorporate both qualitative and quantitative methodologies that may vary between scholarly communities. As Creswell and Clark (2017) describe, some research questions benefit from a combination of both paradigms. Because of this, this research is positioned as post-positivist, with elements of pragmatism. This means that I operationalise the scientific method and understand that there is an

[1]I define "experiments" based on the definition provided by Cohen (1995). He defines experiments as the testing and manipulating hypotheses or causal hypotheses coming from experiments that are manipulated and/or observed

objective world. However, I also understand that the world is multilayered (Cohen et al., 2018), meaning that I acknowledge different interpretations of an objective reality. That is, the objective reality is the basis for an interpretation filtered through various perspectives, which will be seen in Section 2.2.1.

NLP research is traditionally grounded in empirical testing of hypotheses for system evaluation (Cohen, 1995), a trend clearly visible in MT research practices (Marie et al., 2021). Because I adopt pragmatism, I prioritise research questions over methodological purity and accommodate both deductive and inductive reasoning, integrating objective and subjective perspectives (Tashakkori and Teddlie, 1998). This research has to be shared and reported with the assumption that the causality and effects on variables and their manipulation through teaching, careful observation and measurement will bring answers. Accordingly, while this study acknowledges the NLP community's emphasis on deductive rigour, it also embraces the interpretive and reflective engagement afforded by inductive approaches in order to capture the nuanced realities of the studies conducted in this doctoral research.

2.2.1 Foundations of Translation

Defining translation is difficult, as it is best described as a complex process on different dimensions, whether linguistic, social, technological or cognitive (Castilho et al., 2018). In TS, there are several ways to operationalise concepts and accordingly several definitions. That is because TS has a relationship with many disciplines, such as linguistics, comparative literature, cultural studies, philosophy of language, sociology, and history (Munday et al., 2022). That leads to definitions and adaptations that are attached to a specific area of study, such as teaching and, as Munday et al. (2022, p. 22) state, "ever-increasing input from information technology to cover computer-assisted translation". Furthermore, Pym (2023) argues that professional translators are constantly engaged in theorising, whether these theories are formally grounded in a discipline or simply pragmatic ways of solving problems. Pym (2023) highlights the practical benefits of approaching problems without established so-

lutions through theory, since this fosters creative approaches. In this subsection, the goal is to find a theoretical basis to set a foundation that sustains how MTQA is transformed from theory into practice in this thesis and in the course content (Chapter 5).

2.2.1.1 The Goals of Machine Translation

Kuhn (1962) states that understanding the history of a field is essential to grasp the relation between the field's objectives and existing or future theorising. In this thesis, it aids in understanding what is the goal of MT and how the theoretical framework proposed can be aligned with the design of the training proposed in this thesis. Because of this, it is essential to examine the report written by Bar-Hillel (1959), which provides a comprehensive analysis of the developments and challenges from the early stages of the field of MT. He discusses that while progress was occurring, MT development fundamentally required a compromise: it needed to either sacrifice quality for self-sufficient use or it must become an aid to human translators. His words define the extent to which a machine can approximate the act of translation: "Let us be satisfied with a machine output which will ever so often be neither unique nor smooth, which every so often will present the post-editor with a multiplicity of renderings" (Bar-Hillel, 1959, p. 9). According to the author, to reach automatic and full translation MT would need "quasi-human intelligence, knowledge and knowledgeability". This argument was later repeated by Wilks (1978, p. 27) discussing the goal of AI research, which entails the use of computational methods to simulate human behaviour.

2.2.1.2 The Epistemology of Translation and its Assessment

Epistemology is the understanding of how knowledge is created and, thus, its overarching principles extend to how theories and concepts are understood. This is particularly important in the context of this research because it helps operationalise concepts in this thesis, primarily related to MT and to evaluation. First, translation

and by extension MT will be discussed, and secondly, it will be applied to MTQA. I consider "hermeneutics" as the basis of my understanding of translation. For that, I define what is "hermeneutics" first. It is considered either an "art", a "science", or a "theory" that uses interpretation as a basis, developed as a philosophical tradition by Schleiermacher, Dilthey, Heidegger, Gadamer and Ricoeur (Malpas, 2014, p. 2-3). Either as a school of thought or a methodological approach for knowledge construction, it is suggested that hermeneutics must become a "practical philosophy" or "philosophy of praxis" (Vattimo, 2015, p. 723). In other words, its purpose in this thesis is to be operationalised for MTQA.

In TS, hermeneutics has been used as a theoretical foundation by several scholars. The most well-known is Steiner (1998), who sets a theoretical foundation that does not depend solely on purely linguistic and structural analysis and embraces the act of interpretation. This approach was further developed by Blumczyński (2016) and Venuti (2019), emphasising methodological contributions and the refinement of concepts. I will summarise their theoretical elements in seven points:

(1) The translator is an active central figure in translation and evaluation.

(2) Translation is a process that will transform a source text (ST) into a target text (Tt) that will be modified by a human.

(3) Translation requires intervention, as there is no access to the ST without some type of mediation.

(4) The translator subscribes to a hierarchy of values shared (or not) with the institutions or parties interested in the translation or evaluation, in addition to the features and effects of the ST that will be interpreted.

(5) The translator and/or evaluator is an active entity that makes choices based on their expertise, values and ethics.

(6) The process of translation and evaluation provides answers on what constitutes the product and the process.

(7) Although it is possible to attempt a complete prediction of translation, it ultimately needs a medium and/or mediation one way or another.

These seven key points constitute the principles that guide my understanding of MT, its evaluation and other agents[2]. The keypoints will be described in Sections 2.2.1.3, 2.2.1.4, and 2.2.1.5. The theoretical-methodological foundations of Steiner (1998), Blumczyński (2016) and Venuti (2019) will also be contrasted with other translation theory concepts when necessary.

2.2.1.3 Translation and evaluation as interpretation, translators and evaluators as active entities

Key point 1 establishes translators and evaluators as central figures, and **Key point 2** establishes translation and evaluation as a process, to transform and assess the transformation of ST into Tt. These two points are both informed by Steiner (1998), focused on translators for written texts only. Steiner (1998) takes as his theoretical starting point the approach proposed by Nida (1964, p. 126), who cautions that "one must not imagine that the process of translation can avoid a certain degree of interpretation by the translator". In other words, Nida asserts that translation is an active process performed by the human acting as a translator. Further, Nida (1964, p. 127) defines what influences the act of translation:

1. The nature of the message
2. The purpose or the purposes of the author (and by proxy, of the translator)
3. The type of audience

For these three factors, Nida theorises that the translator must make choices and decide the degree to which a translation must follow either the form or the content of the original text. Based on that starting point, Steiner (1998) establishes **the**

[2]I present scenarios where MT systems should be used by translators, and where relevant, by other users. For example, when MT output requires modification and evaluation made by an expert, these principles apply. When MT is used for gisting, where the goal is simply to understand the information for a specific purpose (Bowker and Ciro, 2019), they also apply.

first approach of Translation Hermeneutics (TH). Steiner (1998) expands Nida's thinking to constitute his own operationalisation of interpretation. In this manner, Steiner (1998, p. 29) defines translation as a transformational process from ST to Tt. Any transfer is considered an "interpretive transfer" (Steiner, 1998, p. 29), which, according to the author, it is an active rejection of the analogy of translation functioning as "encoding" and "decoding". This definition accounts for both a process of transformation and interpretive transfer, an important basis at the time to see TH as an epistemological approach within TS. The **second approach to TH** is established by Blumczyński (2016), who draws from philosophy, theology, cognitive linguistics and anthropology. He defines translation not as an object of study, but as a phenomenon or an experience, deviating from an "objective" concept. As such, Blumczyński (2016) defines translation at three levels:

1. An almost mechanical process with a degree of interpretation. This is the level at which machines operate, mimicking human interpretation with pattern matching.

2. A process where most translators operate, with active interpretation. Translators reflect and critically examine texts. They use their knowledge and world experience at this level.

3. A process where translation borders an "art". An epiphany, brilliant idea or a philosophical question becomes part of the process. It may be related to every aspect of the text, whether questioning it or finding a different solution for the translation. It contemplates many elements of the process: ethics, clients, the project, content, and even the translator's personal feelings towards the text.

In this way, referring to these levels, Blumczyński (2016, p. 54) draws parallels to the use of technology. When a translator uses a computer-assisted translation (CAT) tool, the CAT tool provides a translated segment from a database or a translation is generated by an MT system. This is equivalent to the first level described by the author, as the database and/or MT system would mimic the interpretation

by providing a suggestion based on what is likely expected. It is a mechanical comparison of the translation process by providing an already interpreted output, not because there was deliberate choice, but because statistical or pre-programmed rules oversee the mechanical translation process. When a translator accepts the suggestion, rejects it or edits it, this is the second level of interpretation. While the scholar provides no detailed discussion at the third level of interpretation, it is possible to make an inference about it. The translator could be examining the output produced by a CAT tool/MT and reject it, as the suggestion may not be creative enough, might not be critical enough or might not "fit" the text. In this process of examination, the translator could think about the purpose of their work, and then, interpretation goes towards the levels the author described. The deliberate and active choice to think, question or choose words, sentences, or paragraphs requires more depth, creativity and/or critical thinking. This explicit mention of translation technology within theory, albeit brief, opens an opportunity for deeper discussions.

The **third approach to TH** arises from Venuti (2019, p. 7-8), who defines instrumentalism as the "reproduction of a source-text invariant" in translation. In this way, the author positions their model as an interpretive act. Instrumentalism is characterised as a dominant paradigm within translation theory. Here, the ST has an invariant form, meaning or effect (Venuti, 2019, p. 1). The author rejects the notion that the replication of a message through translation is possible, arguing against the definition provided by Nida (1964). In Venuti's definition, translation is an interpretive act that will vary the source-text form, meaning, and effect, by its very nature. That happens according to the interests in the receiving culture, even if it attempts to adhere to strict definitions of equivalence, for example. In other words, even with the strictest rule-based approaches to language, transformation will happen one way or another. Venuti (2019, p. 177) sets forth an invitation to "explore the interpretive power that it commands to change forms and practices, discourses and institutions" when referring to translation. This suggests the inherent transformation that humans are able to perform, which cannot be replicated by MT

without the help of humans.

2.2.1.4 Interpretation as a mediation of values, expertise and ethics

Key point 3 assumes that translation requires intervention, Key point 4 addresses that translators adhere to personal and professional values and Key point 5 establishes translators as active entities when making choices, and they are informed differently in each TH tradition. Steiner (1998) emphasises that translators through their work and theorisation show great empirical evidence and a legacy, philosophically and cognitively, about the process of translation. It is thanks to translators and their translation research that the boundaries of theory and reflections about translation are pushed even further. Steiner (1998) describes every understanding as an active interpretation, because translation concerns the "almost bewildering bias of the human spirit towards freedom" (Steiner, 1998, p. 497). Blumczyński (2016) asserts translation as an experience, and this understanding adheres to how Steiner (1998) perceives his definition of translation as a human process. Translators' jobs always involve ethics, as they constantly interpret and make choices. These choices, for example, can entail refusing or accepting a job on political grounds or using one specific word over another for cultural reasons. They make choices because they think. Thus, translation is not about hiding one's bias, but acknowledging bias, being transparent about decisions and how they impact the translation, people, and the world. When a mechanical interpretation has been made, assigned by a computer, these dimensions are not considered unless explicitly programmed. Venuti (2019)'s approach leads to a process more linked to mediation, through a "succession of interpretations in various forms and practices, media and institutions" (Venuti, 2019, p. 3). The practical criteria for these interpretations are "formal and semantic features specific to the receiving culture and its institutions" (Venuti, 2019, p.176). Those institutions mentioned by him can be defined as professional associations or institutions that uphold laws. In this way, different interpretations, or active choices, can be explained by translators through their expertise and ethical behaviours. This

does not mean that a translator would simply change a text on a whim to sabotage it. While translators mediate and transform texts, the choices behind those interpretations are bound to a "hierarchy of values, beliefs, and representations validated by those institutions" (Venuti, 2019, p. 174), while being wary of the "uncritical acceptance of invariance" (Venuti, 2019, p. 39). Steiner (1998) declares the presence of the translator as an inevitable part of interpretation, Blumczyński (2016) focuses on the translator experiencing and living through the interpretive act when translating, and Venuti (2019) celebrates the act of interpreting as a challenge towards the mechanical conceptualisation of translation. These three perspectives showcase features a human understanding and approach through interpreting. This human understanding translates not just in freedom, but responsibilities and consequences in the process of translation. In the theoretical perspective of TH, the bias of the translator is acknowledged. The ethical, political and social elements of the interpretive acts must be defended. What the translator upholds and believes validates those decisions, while also being informed by their expertise.

2.2.1.5 Translation Hermeneutics and Evaluation

Key points 6 focuses on the process of translation and evaluation through interpretation and Key point 7 specifies that this process cannot be fully predicted. As such, TH addresses the inevitable variability of translation in many regards. Each human has their own way of approaching language, as Steiner (1998, p. 479) declares, each human being "offers its own reading of life". This subjectivity has implications not only for translation but also for its evaluation. Language can involve "concealing" or "leaving unspoken" (Steiner, 1998, p. 48). Whether intentionally or unintentionally, this view of language means that evaluative judgements are shaped by individual interpretation. This positioning shows that humans lack full, mechanical, deterministic prediction when interpreting language and culture. Humans will always act as a filter in the process of interpretation.

The filter that humans possess is not stable: it might not provide consistent

decision-making between interpretations of translation and evaluation. With the idea that interpretation is not stable, I propose that subjectivity interacts with language through layers. These layers may destabilise or stabilise the interpretation. Blumczyński (2016, p. 12-13) states that words or messages are context dependent, adding difficulty to attempts of precise prediction, and words cannot be communicated without a medium. This contributes to a "subjective dimension of meaning formation" (Blumczyński, 2016, p. 10), creating more "layers" that can affect interpretations. For example, Castilho and Knowles (2024) explain how world knowledge can impact translations and linguistic tasks that carry the expectations of what must be accomplished. If a human lacks the necessary world knowledge when evaluating a translation, it will destabilise the human's understanding of it. Or, if the purpose of the task is not understood, it adds a layer of complexity, as the human's interpretation has failed to recognise key elements that must be considered in the process of interpreting. Because of that, context and world knowledge may destabilise or stabilise interpretations, but it is precisely what allows humans to translate and evaluate.

However, the filter of interpretation can also be stabilised. For example, depending on the profile of the human, their ability to interpret the process of transforming or judging an original text into a Tt depends on their expertise with translation. Skilled professionals, when performing a task (for example, an evaluation task judging the output of an MT system), can stabilise that interpretation further through their lived experience and expertise. However, individuals without expertise in translation can have a less stable interpretation when asked to tell how "good", "bad" (or any other predefined criteria) a translation is. There are strategies to be adopted for non-skilled individuals, such as by asking them to interpret a set of instructions to constrain biases. But even with constraining strategies, there is always room for biases, and humans can differ from each other. Knowles and Lo (2024) observe that annotators can be consistent or inconsistent in their own evaluations or among each other in crowdsourcing evaluations, further supporting this (in)stability. This is why

it is likely that when provided with a set of instructions and a task, two translation experts are more likely to achieve agreement in their decisions, but even in such a case with ideal conditions, agreement is not guaranteed. Providing training or instructions about the required constraints of language, culture, and other agreements (such as laws, ethical codes, common agreements in communities) can reduce bias, but are always on the whim of interpretation. And these principles can vary further depending on the type of interpretation. Personal, creative, and literary interpretation might vary more, while technical, scientific interpretations can be more stable. (Doherty et al., 2018)

Since we are filters, we humans must acknowledge the subjectivity inherent in our interpretation of the material world. We must then be held accountable for our interpretations, as expertise and values bind us to the expectation of properly interpreting the transformation from ST into Tt. Because we are humans, we are (un)reliable to different degrees. Humans can be trained to understand or navigate that subjectivity, but it is a human fact that subjectivity may be unintentionally hidden or cause people to not notice what has been asked.

In practical words, an ideal or acceptable translation will be achieved by a translator with their interpretation. That is also expected to happen for evaluations made by translators. A non-expert, on the other hand, will provide different interpretations and it might be difficult to understand their choices. As Venuti (2019, p. 3) states, translation evaluations and their thoroughness occur according to their importance and expectations. This means that the needs guide the evaluation. That suggests that not only ST features and effects guide an evaluation, but also the values, beliefs and representations upheld by parties for whom the translation was created. Thus, for points 6 and 7, I suggest that TH addresses the weaknesses of HE by exposing the subjectivity and the bias involved. But it also cements the path to add accountability through constraints and acknowledging the human behind the decisions. In essence, TH sets a theoretical foundation that accommodates the weaknesses of subjectivity with practical ways to remedy it, making translation

and evaluation more stable.

2.2.1.6 Translation Hermeneutics and other translation theories

Examining TH in comparison to other translation theories is useful to correctly position TH and understand its boundaries, mainly to verify TH's alignment with MTQA as an activity. Further, the examination of other theories and perspectives is also an element of TH itself. Steiner (1998, p. 129) mentions that every statement about language or translation "invites a counter-statement or antithesis", as discussions of language or translation can be "unstable" and "dialectical". For this reason, it is productive to contrast other theoretical approaches from TS with TH, such as the concepts of equivalence informed by Nida (1964) and functionalism informed by Nord (2005, 2014). The purpose of this comparison is to discuss where they may differ or coincide in different aspects. This contrast is not intended to be exhaustive, as these theories are not necessarily adverse to TH, but are operationalised in different ways. TH can contrast or coincide with the concept of equivalence, because TH aims to focus on translation as a process of transformation. For example, formal equivalence conceptualises Tt to "match" as closely as possible with the ST (Nida, 1964, p. 129). On the other hand, dynamic equivalence, where there is less emphasis on reproduction of the form, but more emphasis on reproduction of the "same message" in the Tt (Nida, 1964, p. 129), transformation can happen. The key difference is that TH suggests the explicit intervention of the translator. That intervention happens through the translator's active decision to translate using their interpretation, while of course not forgetting about the interests and needs of the parties who are using the translation.

Equivalence assumes either the formal aspects of language or the reproduction of language functions. It considers the ST as the main factor in the process (Pym, 2023, p. 39). It shares aspects with TH on the creation of criteria, particularly when setting criteria for the purpose of each translation. However, it differs from TH on grounds that some types of equivalence focuses solely on the ST itself. The focus

on ST distances it from TH, because TH also accounts for:

a The process of translating

b The Tt

c The Translator

d The stakeholders interested in the translation (such as clients or the users of the translation)

e The inevitable transformation, variation or mediation that the translator performs.

In some respects, Nord (2014, p. 19) presents translation as an "intentional interaction", on the part of the translator who will follow the original sender or text producer in the process of translating a ST. According to this theory, the translator "produces signs", the translator is responsible for "chosen informational items" that are supposed to be "transferred to a target culture", if the translator believes it is appropriate for a purpose given. Even if there are interpretive acts carried out by the translator in the process, Nord (2005, p. 5) explains that the emphasis on function should be "the overriding criterion for the model of the translation process". That is, it is a compromise that the function may supersede any other aspects of the process. In a way, this could limit the translator who subscribes to functionalism. Translation as understood under TH, on the other hand, will vary the ST form, meaning and effect. It considers values, beliefs and representations. It offers more flexibility, as the translator is not and should not be a neutral party, when examining the position of the translator in terms of translation theory.

2.2.1.7 Translation Hermeneutics and relations to human or machine translation

This subsection will be used to examine MT in different aspects and how definitions resonate with TH. In examining how MT is positioned in TS, Asscher (2023)

concludes that most conceptual definitions of translation within TS work well with how MT functions, specifically NMT. For example, Catford (1965) operationalises translation in two concepts, where translation is the semantic mapping from one language to another and transferral is a transference of named entities. Conclusions from an experiment involving NMT in the domain of literary translation (Moorkens, Toral, Castilho and Way, 2018) draw parallels between the machine learning method and the concept of transfer from Catford (1965), stating that it is problematic to compare a process of transfer with "attempts to replicate the patterns in training data" (Moorkens, Toral, Castilho and Way, 2018, p. 256-257). Moorkens, Toral, Castilho and Way (2018)'s results demonstrate that the lack of contextual information beyond the linguistic side is a major weakness from MT systems. Those results are useful, because it resonates with Asscher (2023)'s conclusion that translation conceptualisations that attempt to differentiate HT from MT are useful, particularly in respect of "conscious agency" as to highlight the main difference between human and machine. This resonates with whatMoorkens, Toral, Castilho and Way (2018, p. 256) conclude, framing the machine learning method as "non-conscious algorithm[s]", differentiating it from human processes of translation and decision-making.

To discuss MT, it must be compared to HT, as we need to reflect how both constitute relevant parts of the translation process. Outside TH, the definition of HT in comparison to MT can be drawn from Kenny (2018, p. 424), who defines HT as a process, carried out by a human, who converts a ST into a Tt sometimes with the aid of translation technologies. According to this definition, HT incorporates mediation, the human acts as the primary agent. When comparing the definition of HT to how MT operates, the machine requires the aid of a human, who is still the primary agent. In the case of LLMs, prompting can be the intentional guiding of human input. This aid can happen before or after carrying out the interlingual conversion of text (Kenny, 2018, p. 429), for both type of systems. Another definition that highlights the position of the machine comes from Asscher (2023, p. 6) who describes MT as

both directional and subjectless. It is reminiscent of formal equivalence (Nida, 1964, p. 129), alluding to the lack of a human during the process.

Bringing these definitions to TH, I can examine the assumptions of how translation is performed by humans in comparison to machines. Consider human data used, which includes their interpretive acts, decisions and mediations. Or consider humans who have used their interpretive actions to evaluate or edit the output of MT systems. Machines are specified as the "tool" during the process. Although it can be argued that if the human decides, MT as a "tool" can become a "primary agent", MT is still devoid of unique interpretations and only produces the interpretations on which it was trained or which it is ordered to.

From the perspective of computer science: "Translation is always an approximation. Translators have to make choices, and different translators make different choices" (Koehn, 2020, p. 4). Even though this perception comes from outside TS, MT systems are are built efforts of developers and translators, sometimes a joint effort, sometimes not. It is natural that computer scientists have their own perspectives on how the process happens, Koehn views MT's aim as conversion through approximation. In my view, it is not a process of interpretation as an act, rather, it is the computation of interpreted acts carried out by the machine through the method of prediction. As Moorkens (2024, p. 10) discusses philosophically, any translation prediction would be based on material contained in the training data, complete with its biases and no opportunity to update it without more data.

When MT is used without the mediation of a professional human translator, however, it introduces the risk of producing unwanted or misleading output for non-expert users. Unlike human translators, MT systems do not possess inherent safety filters or contextual awareness. Large Language Models (LLMs) represent a partial exception, as carefully designed prompts can serve as limited safety mechanisms. Nevertheless, the process remains unmediated by an expert capable of ensuring interpretive and ethical safety.

2.2.1.8 Machine Translation Quality Assessment

As mentioned in section 2.2.1.2, humans serve as filters with more or less stability depending on their professional profile when evaluating translations. The concept of quality itself can be interpreted through different dimensions (Castilho et al., 2018). House (2014, p. 1) states that the concept of quality assessment is at the heart of any translation theory, that is, a translation theory is also a theorisation about the concept of quality. The author emphasises the linguistic aspect as a core principle for the operation of translating, and as a result, for measuring quality. This coincides with Venuti (2019)'s understanding that the interpretation must be based not solely on values and expertise, but the linguistic content. Not meaning that other theories disregard the linguistic content, but very clearly they possess different degree of values between theorists.

TRUEFor House, translation is the involvement of "the translator's knowledge, expertise, ethical stance and attitudinal profiles as well as her subjective theory of translation" (p.3). When comparing that view to translation under TH, it is in fact to accommodate such a perspective, even if House (2014, p. 8-10) condemns a hermeneutic perspective of translation, because it focuses only on subjectivity. While it is true that the hermeneutics values subjectivity over any material reality, I argue that my conceptualisation of TH considers the interpretation of a material reality. That is, it is based on language (material), but has the openness of various interpretations (subjectivity), making it inductive in a post-positivist approach. In the context of this thesis, this means that I construct knowledge in this same orientation, by observing a reality by analysing my object of research and applying my own interpretations to describe what I am observing, much like how a post-positivist orientation to research may function.

Castilho et al. (2018) explain that approaches and practices vary in the language industry. The increasing interest in MT systems from a research or a commercial point of view has contributed to different ways of checking quality and different quality standards, depending on the purpose of the MT system and the tasks related

to its evaluation (Castilho et al., 2018, p. 11). HE's weakness is commonly attributed to subjectivity, as it can be quite unstable. AEMs also have their own weaknesses, with important distinction between approaches. Reference-based metrics compare MT outputs against one or more translations produced by humans (considered the "gold standard"). While useful, they emphasise similarity and risk neglecting the variability inherent in translations produced by humans (Way, 2018). Reference-free metrics, on the contrary, do not rely on gold standards but instead use models trained on large amounts of HE data. These models aim to approximate human judgements, as they are trained by learning from patterns of HE metrics or specific parameters and features set by developers. However, because there is a human element through the annotations, they may replicate biases. In addition, they may make calculations based on spurious data and not necessarily semantic meaning (Perrella et al., 2024). These factors must be considered when interpreting the results of AEM scores, and show that the human behind the design of the metric and its choices can render the scores themselves more or less stable, as discussed previously. This means that evaluation is unstable by nature. Recognising this subjectivity of interpretation, layers of stability can be added with tools and strategies present in MTQA research, such as seen in Figure 2.1:

(a) Guidelines

(b) Evaluator Expertise

(c) Evaluation platforms

(d) Likert Scales

(e) Test suites

(f) Risk assessment measures

(g) Methodologies validated by empirical research

(h) Training

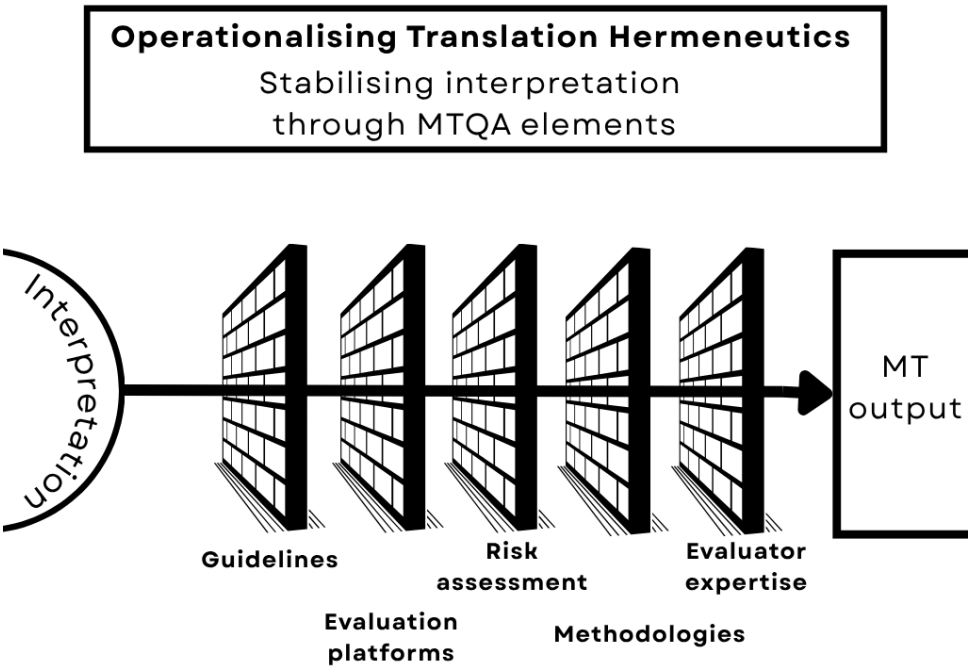


Figure 2.1: Illustration representing how interpretation attempts to penetrate through constraints represented as solid walls. Each solid walls represents layers that stabilise interpretations as it passes through each wall, in order to reach the MT output.

ATRUEll of these tools and approaches may stabilise interpretation in MTQA. As suggested earlier, interpretations can be stabilised by passing through layers. For example, guidelines are instructions that carefully lead the interpretation of evaluators towards the desired type of interpretation, which must be respected according to specifications of a quality assessment. Guidelines have been shown to impact annotator's evaluations, as their interpretation of the guidelines affect the interpretation of the MT output (Knowles and Lo, 2024). Evaluation platforms contain elements that can direct workflows or attention towards specific elements (words, sentences, text boxes, options, sensitivity of sliders) that affect the experience of the evaluator and may affect or constrain the interpretation of the user of that platform (Knowles and Lo, 2024). Risk assessment measures are constraints that direct interpretations to focus on certain elements or features that may be present in the MT output. They are taken to prevent interpretations that could cause harmful consequences, such as financial and reputation risks (Canfora and Ottmann, 2018). Research-

based methodologies are recommendations made by experiments that have tested constraints in MTQA and reported results quantitatively and/or qualitatively. This renders the recommendations as means to stabilise interpretations, and as a consequence, to stabilise results. This is why the recommendations such as ones made by Läubli et al. (2020) and Marie et al. (2021) provide means to focus on specific effects of MTQA design to achieve more stabilised interpretations. Evaluator expertise is one example of stabiliser mentioned previously, which can be enhanced with descriptors in international standards that define conditions for evaluators that have the knowledge to stabilise interpretations. Training is particularly effective because it provides that expertise.

Figure 2.1 shows these constraints as walls. They are not impenetrable. They add solid, material, rigid conditions that impact how an interpretation is made given an MT output. Like real walls, each brick (or feature) is selected by a human, and may prove to be effective or not. Thus, even those barriers can differ in effectiveness depending on how well each brick is placed, to direct the interpretation towards the desired (stabilised) outcome. These principles must be understood in MTQA design. Solid understanding of the philosophical process of TH guides the pedagogy of the proposed training course of this thesis.

2.2.2 Foundations of Education and Instructional Design

The educational concepts that operationalise our pedagogical framework are found in this section. Concepts from education influence how the curriculum, lesson plans, and practice in the classroom are performed, while concepts from ID guide the overarching framework that provides the methodological phases for this doctoral thesis. They will impact the description of methods (Chapter 4), the design of the training (Chapter 5) and the implementation of the training, with emphasis on the classroom experience (Chapter 6).

2.2.2.1 From epistemology to curriculum and practice - Freirean Philosophy

Defining curriculum means examining the epistemological, conceptual, and ideological views of the educator (Connelly and Xu, 2010). For this thesis, Freire (1996) provides a compatible pedagogical philosophy with my translation principles and the overarching goal of this thesis, that is, to create MTQA training for master's students in NLP. All of the principles from Pedagogy of Freedom Freire (1996) are useful and constitute my practice as a researcher and educator. However, one of the recommendations Freire (1996) makes is that educators should extract principles from philosophy for active reflection. Because of that, the principles I considered relevant and compatible with my practice were extracted and highlighted for application in this doctoral research. Their main role involves guiding and justifying my pedagogical decisions throughout the training. Doing so will produce knowledge to create a teaching methodology that centres humans when interacting with MTQA and its elements. The chosen principles were the following, drawn from Freire (1996).

(i) Teaching requires methodical rigour and research

(ii) Teaching requires critical reflection about its practice

(iii) Teaching requires respect towards the autonomy of students

(iv) Teaching requires conscious decisions

In (i), Freire (1996) uses the concept of rigour and research to explain that learning is an active process, and therefore, it is understood that teaching is not a transfer of knowledge. Students are not empty, passive vessels. Teaching students involves leading them to engage with the content and the ability to think by themselves. The student must have the conditions to interpret how they interact with the world critically, interpreting the content with their own background knowledge to make meaning by themselves. It is through research that a teacher can create the conditions for students to see and interpret the world with a critical lens. That means

the teacher will foster a sense of curiosity, urging the students to be protagonists in their process of learning, so they feel the need to create and engage inside and outside the classroom.

In (ii), it is understood that teaching involves a dynamic movement between performing an action and reflecting about it. Freire (1996) affirms that the theory must be so concrete and practical that it must be almost mistaken for practice. That is, the goal is to analyse the object of study as closely as possible through the use of theory. Because it is a human process, part of reflection also involves emotions. It is not possible to distance oneself from reflection. Freire (1996) affirms that part of this process involves feeling emotional, to feel anger against injustice, against unfairness, against exploitation, against violence, all of these feelings are part of learning. Learning is also a personal and human process. Reflecting is to be human and act human.

In (iii), it is understood that teaching is not simply the interaction of student and curricular content, but associating the curricular content with the experience students may bring. Freire (1996) suggests that critical pedagogy involves combining the curricular content with the reality of the world, questioning the political and ideological decisions that impact people. Teaching requires students to examine the curricular content with an ethical lens, investigating how that content exists in the world. This is how students examine the curricular content and learn their own thought processes, with their own background, aware of how the world is and how it can be improved.

In (iv), it is understood that education is an act of intervening in the world. It is through this principle that the teacher acts in the most humane way when speaking to students. That means addressing their dreams, ideals, utopias, and objectives, because these elements impact the world. Education is political because the curricular content is not neutral. Naturally, as humans, we embed our own interests in our teaching, whether that is an optimistic interest or pessimistic interest. As Freire (1996, p. 127) states that "scientific and technological progress that does

not correspond fundamentally to human interests, and to the need of our existence, lose, to me, their meaning", a teacher must position themselves when teaching about technology, especially when teaching students who develop technology and work on improving it. The duty of a teacher is not to stop research or technological advances, but to highlight to students that research and technology must be in the service of human beings (Freire, 1996). Students are part of a system that may not allow them to fully enact change, but so long as they are aware of their actions (or inaction) in the world, education can be critical.

These four principles are the most compatible with TH in my view. As interpretation is an active process that is affected by values, ethics, and expertise, it makes sense to provide a learning experience that allows students to focus on practising these principles while adopting a Freirean lens. That is, students can interpret the world and make active decisions to develop technology in a way that can potentially change how they benefit humanity. Furthermore, the purpose of Freirean pedagogy is to make students critical by comparing, repeating, affirming, questioning, and being curious to highlight these actions that show their value as humans. Furthermore, these principles are also my guiding direction in assessing my own practice, being useful not just in the teaching of the training course but also in reflection. Through this philosophy, I reflect on the students' interactions, assignments, and feedback, so I can link practice to theory.

2.2.2.2 Instructional Design

For the same reason why the history of MT was explored in Section 2.2.1.1 based on the importance of historicity for epistemology (Kuhn, 1962), it is important to clarify how research in curriculum development transformed over the years, exploring General Didactics (GD) and ID. The purpose of ID is to create a learning environment by choosing the appropriate pedagogical theories to effect interventions and evaluate them (Seel et al., 2017). The field of GD serves very similar purposes and has had similar goals to ID (Zierer and Seel, 2012). According to Zierer and Seel

(2012), GD is focused on the relationship between teacher and student, teacher and content, and student and content, which can be structured in different models. As scholarship in teaching grows, more definitions have arisen. A bibliometric study has shown taxonomies such as "learning design" or "learning engineering" (Saçak et al., 2022). As Saçak et al. (2022) explains, learning design is focused on the learner's experience, differing from the content-centric approach of ID, while learning engineering has emerged as a need to focus on scientific and evidence-based studies in learning.

Whatever the differences, the focal point to be understood here follows the assertion about models and frameworks, either from ID or GD, made by Zierer and Seel (2012, p. 9): "They are designed to help solve the main educational problem of theory and practice, to build a bridge between and overcome the specific differences between theory and practice". This can be applied by the taxonomies denominated by Saçak et al. (2022). There is a danger of model fatigue as Zierer and Seel (2012) describe, when there is a need to exhaustively create models to advance learning science. Furthermore, Saçak et al. (2022) believe that ID and learning design can be used interchangeably, and that the future direction should be the validation of epistemological frameworks with certain methodological choices to successfully validate the uses of ID in research and learning. This is supported by Zierer and Seel (2012) explaining that the future of didactics should be the incorporation of the critical aspect, hence my use of Freire (1996) with my intervention.

The frameworks adopted from ID serve to plan and analyse the central educational problem of this thesis. Since ADDIE is a general model (Zierer and Seel, 2012), its whole purpose is to structure the phases of the investigation. Furthermore, there is one practice from GD that benefits this project, which is the concept of eclectic didactics. It is the process of having a good understanding of a learning theory and integrating additional components into the process as necessary (Honebein and Sink, 2012). Additionally, because the SRL in Chapter 3 will involve various teaching approaches, being eclectic is essential to absorb the most appropriate teaching

practices to be embedded in the training course proposed in this thesis. Thus, this justifies the pedagogical rationale choices made.

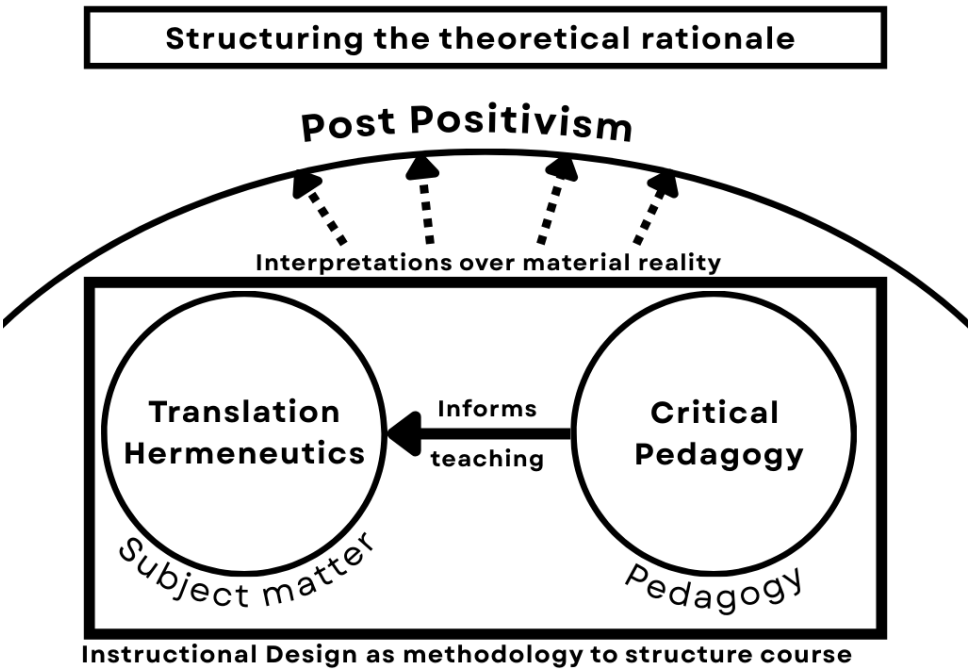


Figure 2.2: Theoretical rationale informed by post-positivism, aligning translation hermeneutics and critical pedagogy in dialogue, with instructional design as the structuring methodology for teaching practice.

Figure 2.2 serves to summarise how all the theoretical components compose the rationale for this course. The combination of critical pedagogy and translation in the perspective of TH will be embedded in the curriculum. Using ID as the overarching methodology, it is through this combination that classroom data will provide interpretations that attempt to explain reality.

Chapter 3

Systematic Review of Literature - "Needs Analysis"

This chapter corresponds with Phase 1a of this doctoral research, centred on addressing Research Question 2a: "What research has been published on Machine Translation Quality Assessment teaching for Translation Studies and Natural Language Processing?". It is dedicated to exploring the teaching of MTQA, with an emphasis on HE, with the goal of identifying practices in TS that can be integrated into NLP education. The initial stage of the ADDIE framework (Seel et al., 2017), focuses on needs analysis. It serves the purpose of: i) discerning the future learning needs and performance expectations of the learner; ii) assessing prior educational experiences; iii) defining the scope of future educational content (Seel et al., 2017, p. 70). To achieve this, a Systematic Review of Literature (SRL) has been conducted, guided by principles articulated by Petticrew and Roberts (2008), Gough et al. (2017) and Zawacki-Richter et al. (2020).

This literature review method was selected due to the extensive teaching tradition of quality assessment in TS, encompassing both human translation and MT. Only publications about MT will be selected, as studies related to human translation are not relevant to the scope of this research. A systematic approach is well-suited for identifying studies that can inform educational practices (Newman

and Gough, 2020) and highlight potential research gaps (Petticrew and Roberts, 2008). In essence, it facilitates the retrieval of educational resources, rationales, and curricular choices from TS regarding MTQA, which may possibly be adapted for NLP education. Furthermore, it aids in recognising potential gaps in the literature, particularly within the scope of this research focusing on the teaching of MTQA to NLP audiences.

To demonstrate the process of the SRL from start to finish, the methodology employed for the SRL encompasses the following sections: a) description of the method (Section 3.1); b) description of the process and conditions for the literature search (Sections 3.2 and 3.3); c) description and coding of data (Section - 3.4); d) synthesis and evaluation of data (Section 3.5). From this process, two other sections are added to this chapter, an application of the synthesis of recommendations to follow to adopt HE in NLP education (Section - 3.6) and a curricular analysis to complement the findings of the SRL (Section - 3.7).

3.1 What is a systematic review of literature?

The purpose of an SRL is to systematically and accountably review research findings. SRLs are categorised into two types: aggregative SRL, which aims to approach specific research questions, and configurative SRL, which employs qualitative approaches to explore topics or examine researchers' and participants' experiences (Gough et al., 2017). Embracing this approach allows a nuanced and iterative understanding of the selected studies in TS and NLP. While naturally aligning with the depth and exploration characteristic of configurative SRLs, I also intend to integrate structured elements typical of aggregative SRLs as I aim to address research questions. In that way, I can engage with my research questions and explore the relevant pedagogical topics.

I embrace a mixed-methods approach that integrates both qualitative and quantitative methodologies and data. This approach enables the researcher to combine data collection and analysis methods aimed at addressing specific research ques-

tions while remaining aligned with their philosophical beliefs and theories (Creswell and Clark, 2017, p. 38). Using a mixed-method approach allows the exploration of questions that might remain unanswered when using either paradigm independently (Creswell and Clark, 2017). Consequently, the research follows a deductive approach, organising data from a broad perspective, followed by an inductive approach, where specific data characteristics are observed to establish principles (Morgan, 2007).

3.2 Systematic Review of Literature Protocol

In order to conduct such a review, the initial step involves drafting a protocol that outlines the questions that the SRL will answer. The complete protocol is available in Appendix A, which provides a summary of the procedures, conditions given to contextualise the literature search, the strings used to retrieve studies, and inclusion/exclusion criteria. This SRL is divided into two stages, made possible due to the exploratory nature of this SRL, which allows flexibility and/or changes in the process (Newman and Gough, 2020), including setting specified time periods, which is a common practice in SRLs (Gough et al., 2017) for individual researchers conducting SRLs. Stage one aims to scope the literature on MTQA teaching to identify and analyse teaching materials, pedagogical procedures, metrics, methods, curricula and syllabi within the field of TS. In stage two, the focus shifts to identifying studies related to MTQA teaching within the field of NLP. It has the following SRL research questions (SRL-Q):

- **SRL-Q1** - How is MTQA taught in TS education?
- **SRL-Q2** - How is MTQA taught in NLP education?
- **SRL-Q3** - How is MTQA taught in different teaching modalities in TS and NLP education?
- **SRL-Q4** - What are the rationales for MTQA in TS and NLP education?

- **SRL-Q5** - What educational materials are used for MTQA teaching in TS and NLP education?

Collectively, these SRL-Qs guide the whole process and aid in answering RQ2a. Specifically within the SRL, each of the SRL-Qs is used to map and classify data into different themes, a process which is described in more detail in data codification (Section 3.4). Furthermore, a series of keywords were chosen to search for the studies (available in Appendix A), in three languages: English, Portuguese and Spanish. These languages were selected according to my language competence, and because studies from different countries and regions may provide additional diversity in data, mitigating English language bias (Gough et al., 2017).

Twelve databases were selected to be included in the study: 1) Institute of Education Sciences; 2) Hispanic American Periodical Index (HAPI); 3) JSTOR); 4) MLA International Bibliography; 5) Science Direct; 6) Taylor and Francis Online; 7) Web of Science; 8) Wiley Online Library; 9) Portal de Periódicos da Capes; 10) Scientific Electronic Library Online; 11) SAGE Journals; 12) Google Scholar.

These databases were selected because they represent :

a) a concentration of studies

b) diversity in fields, including databases that cover a broad range of subjects or ones that focus on a particular field, such as computer science

c) databases from different regions and languages such as Portal Periódicos da Capes which is a Brazilian database or the HAPI database, which provides data in Spanish.

The strings were created with the following criteria:

a) use the keywords set in the protocol

b) use the functions specific to each database, considering their own particular commands and characteristics

c) combine the most relevant keywords to retrieve as many results as possible.

These criteria were established because "it may be necessary to undertake several searches or to use a mixture of broad and narrow terms" (Gough et al., 2017, p. 111). After adapting the string to each database, the total number of entries retrieved from all databases was 4,606 studies: 3,730 studies in the first stage (TS) and 876 studies in the second stage (NLP).

To filter the 4,606 studies found, criteria were established for the inclusion, exclusion, selection and evaluation of studies and are shown in Appendix A. In both stages 1 and 2, studies were screened in a systematic order of examination, commencing with a review of the title, abstract, and keywords to assess their relevance to the SRL-Q. Subsequently, the methodology of each study was examined to gain insights into the study's context and whether they met the inclusion criteria. This procedure is particularly critical, as both stages aim to retrieve educational studies with a pedagogical and methodological focus on teaching. Specifically, these studies were required to emphasise the description of pedagogical procedures or provide pedagogical contributions derived from experimental outcomes. Following the initial screening, the studies were systematically organised, extracted, and presented according to the guidelines outlined in the protocol.

In this process, as per Table 3.1, data was initially generalised to identify themes through deductive reasoning. Subsequently, the data underwent analysis to interpret specific contextual aspects and formulate principles through inductive reasoning. Subsequently, the data is interpreted to gain insights into how MTQA has been taught and to identify the unique contributions of each study, adhering to the criteria established by the researcher.

Strategy	Details
Data Synthesis and Presentation	Data will be organized by means of a description, in forms with bibliographic information (author, title, year, etc.). These forms will help organize the data in order to elaborate a metanarrative.
Data Extraction Strategy	Data will be extracted following a specific strategy, detailing bibliographic information such as author, title, year, etc.
Data Summarisation Strategy	Data will be summarised as a thesis chapter and a paper, with descriptions and analyses of interpretive nature.
Publication Strategy	Results will be written comprehensively as a thesis chapter and will be later written as a paper to be submitted for peer-review.

Table 3.1: Data Synthesis, Extraction, Summarisation, and Publication Strategy

3.3 Data Collection Methods

Two software programs were used to collect the data:

- JabRef: an open-source citation and reference manager. It was chosen due to its user-friendly interface and my familiarity with the program. [1]
- Zotero: an open-source reference manager that was used as a complement, as sometimes JabRef was not able to handle certain files. [2]

To ensure a comprehensive search for relevant studies, multiple databases were searched using specific strings. Appendix A provides a summary of stages 1 and 2 of the SRL, showing the databases, strings included, and results obtained, and quantity of included/rejected studies. Gough et al. (2017) emphasise the importance of

[1] Available on: `https://www.jabref.org/`
[2] Available on: `https://www.zotero.org/`

Stages	Total Studies	First Round Selected Studies (Removing Duplicates)	Second Round Selected Studies (Inclusion and Exclusion Criteria)
Stage 1 - Translation Studies	3730 studies	134 studies	32 studies
Stage 2 - Natural Language Processing	876 studies	6 studies	3 studies

Table 3.2: Number of studies for each Stage of the SRL, including the total number, followed by the removal of duplicates (first round), and the application of inclusion and exclusion criteria (second round).

employing a well-structured search strategy to retrieve as many studies as possible efficiently. Therefore, meticulous attention was given to registering all terms, strings, sources, search limitations, and dates with comprehensive detail. This process encompassed the documentation of results from all databases, along with the specific search strings used. Notably, certain databases, such as JSTOR and ScienceDirect, presented reformatted versions of search strings in their systems. The record-keeping of how each string was displayed across different databases facilitates the aspect of replicability, alongside the inclusion of the date of access, which is essential as it may undergo changes over time. Results can be see in Table 3.2.

In the first stage, 3730 results were found, from which 134 studies were chosen for the field of TS, after selecting likely studies following the strategy of reading the title, abstract, keywords and methods, in addition to removing duplicates. In the inclusion criteria and evaluation rounds (as seen in Appendix A), 32 studies were selected. In the second stage, encompassing NLP, a total 876 studies were found, from which six studies were chosen, after removing duplicates. In the inclusion, exclusion and evaluation rounds, three studies were selected. There was a notable decrease in studies for NLP in comparison to TS, during the screening and the inclusion/exclusion process. In total, 35 studies (32 in stage 1, and 3 in stage 2)

were codified and organised, and are described in detail in the following section.

3.4 Data Coding

Following the data collection stage in the SRL, the subsequent step involves the identification of themes through coding (Cohen et al., 2018). Given the formulation of various SRL-Qs (as detailed in Section 3.2), the studies were systematically coded to align with the respective questions, directly corresponding to the overarching objectives of this thesis. As the primary goal of RQ2a is to discern how MTQA is taught in TS to gather resources and inform teaching practices in MTQA with an emphasis on HE for NLP, the design of the SRL-Qs effectively addresses RQ2a. As highlighted earlier, conducting a needs analysis requires an examination of historical and current educational experiences to inform future pedagogical recommendations. Consequently, the codification process was executed to systematically classify the pedagogical experiences into distinct themes.

The themes listed in Table 3.3 namely Teaching Procedures and Methods (TPM), Curriculum and Syllabus (CSL), Modality (MOD), Rationale (RTL) and Resources (RE), were employed to construct codes that corresponded to each of the SRL-Qs. TPM and CSL are two distinct themes derived from SRL-Q1 to account for varying levels of granularity in addressing the specific focus of individual educational instances (classroom procedures and curricula).

TPM, which addresses experiences at the micro-level, is used for studies detailing procedures for classroom activities, such as reflective tasks in labs related to the use of MTQA metrics and procedures. In contrast, CSL, which addresses experiences at the macro-level, is applicable to studies that describe or discuss decisions related to syllabi or programmes, such as the choice of content. While TPM and CSL themes primarily derive from SRL-Q1, they are equally applicable to SRL-Q2. The distinguishing factor rests in the specific field of study under examination, which is NLP. Intriguingly, within the realm of NLP, no studies have been identified that employ the TPM theme. MOD was reserved for studies that offered insight into

educational experiences conducted in an online or hybrid format in TS. Meanwhile, RTL was reserved for studies that conducted experiments or provided results that elicited insights aimed at enhancing MTQA education. It is noteworthy that RTL studies also featured teaching procedures to some extent or possessed a pedagogical focus on the experiments, even if this aspect was relatively minimal in comparison to the pedagogical themes.

Question	Theme
SRL-Q1 - How is MTQA taught in TS education?	Teaching Procedures and Methods (TPM) Justification: Methodological contribution at micro level (classroom)
	Curriculum and Syllabus (CSL) Justification: Methodological contribution at macro level (syllabi or undergraduate/postgraduate programmes)
SRL-Q2 - How is MTQA taught in NLP education?	Curriculum and Syllabus (CSL) Justification: Methodological contribution at macro level (syllabi or structured programmes)
SRL-Q3 - How is MTQA taught in different teaching modalities in TS and NLP education?	Modality (MOD) Justification: Focus on modality of instruction
SRL-Q4 - What are the rationales for MTQA in TS and NLP education?	Rationale (RTL) Justification: At least minimal methodological contributions, but also contributions from empirical data by the means of suggestions, reflections, guidelines.

Question	Theme
SRL-Q5 - What educational materials are used for MTQ teaching in TS and NLP education?	Resources (RE) Justification: Materials that can be used for teaching, whether as a set of activities, a textbook or resources available online.

Table 3.3: Description of SRL-Questions matched with themes and justification for the topics that must be discussed in each theme

3.5 Data Synthesis

3.5.1 Stage 1 - Translation Studies

This section will focus on synthesising the data of the SRL, focusing on the TS field. To conduct the process of synthesis, the final 32 studies[3] (Table 3.2) were grouped in the different themes (Table 3.3). Table 3.4 provides the total number of studies attributed to each theme.

Theme	Studies	Percentage of total studies
Rationale (RTL)	21 entries	65.63%
Resources (RE)	6 entries	18.75%
Modality (MOD)	6 entries	18.75%
Teaching Procedures and Methods (TPM)	9 entries	28.13%
Curriculum and Syllabus (CSL)	10 entries	31.25%
Total	32 entries	100%

Table 3.4: Total of studies attributed to themes in Stage 1

As can be seen from table 3.4, a substantial proportion of the studies (65% of the

[3]Appendix B shows a full list of the studies which were attributed themes, with studies ordered chronologically and referenced by a number. The studies in each line were associated with one or more themes depending on their contributions.

total), prioritise RTL themes, contributing to pedagogical insights by conducting experiments in various educational and non-educational contexts, and offering guiding principles related to MTQA, including topics, procedures, metrics, and guidance for MTQA pedagogy. Furthermore, TPM and CSL studies (28% and 31% respectively, 59% of the total), represent an equivalent emphasis on elucidating classroom-based experiments and specifying procedures for teaching MTQA or curricular guidance, while providing practical recommendations for implementation, either through classroom practises or selection of content for syllabi. Finally, while the number of RE and MOD studies, each 18%, is not as extensive as that of RTL, TPM or CSL, they offer insights into some effort of sharing resources for MTQA, reporting an implementation of hybrid and online approaches towards the teaching of MTQA, and contribute to the broader MTQA scholarship by providing resources and investigating contextual factors.

Each study in Table 3.5 is referenced by a number, and individual rows are associated with distinct themes or a mix of them. A comprehensive list of studies, correlating each number to the respective author's name and theme can be found in Appendix B.

Lines	Reference Number (REF)	RTL	RE	MOD	TPM	CSL
1	[7], [8], [10], [13], [15], [20], [21], [24], [25], [26], [27], [28], [30]	x	-	-	-	-
2	[3], [5], [6], [30],		x			
3	[6], [20], [21], [30]			x		
4	[6], [18], [22], [28], [29]				x	
5	[4], [6], [20], [21], [28]					x
6	[31]	x	x			
7	[2]	x		x		
8	[16], [19]	x			x	
9	[1], [11], [12], [14], [17]	x				x

Table 3.5 continued from previous page

Lines	Reference Number (REF)	RTL	RE	MOD	TPM	CSL
10	[23]		x		x	
11	[32]			x	x	
12	[9]				x	x

Table 3.5: Studies classified by thematic focus. An "x" marks the presence of a theme within each study: Rationale (RTL), Resources (RE), Modality (MOD), Teaching Procedures and Methodologies (TPM), and Curriculum and Syllabus (CSL). Lines 1–5 indicate single-theme studies, while lines 6–12 represent multi-themed studies addressing two or more categories.

(a) Rationale (RTL):Nine studies addressed RTL (line 1), complemented by 13 studies incorporating RTL components (multithemed studies from lines 6, 7, 8 and 9; studies with more than two themes underlined in line 1).

(b) Resources (RE): Two studies primarily centred on RE (line 2), with an additional four studies presenting RE components (multithemed studies from lines 6 and 10; studies with more than two themes underlined in line 2).

(c) Modality (MOD): No study focused entirely on modality (line 3), only multithemed studies were found (lines 7, 11; studies with more than two themes underlined in line 3).

(d) Teaching Procedures and Methodologies (TPM): Three studies primarily investigated TPM (line 4), with six more integrating TPM elements (multithemed studies from lines 8, 10 and 12; studies with more than two themes underlined in line 4).

(e) Curriculum and Syllabus (CSL): Lastly, one study was centred on CSL (line 5), supplemented by nine studies integrating CSL components (multithemes studies from line 9; studies with more than two themes underlined in line 5).

3.5.1.1 Rationale (RTL)

This subsection aims to identify the prevalent subthemes and topics of studies under the theme RTL, which answers the SRL-Q4 What are the rationales for MTQA in TS and NLP education?. Studies that have minimal methodological contributions on pedagogy, but have valuable contributions from empirical data by the means of suggestions, reflections, and/or guidelines are included. This will guide this research in terms of providing a comprehensive understanding of the importance of HE in MTQA and the position of stakeholders being literate in evaluation processes in order to inform the design decisions for the training for NLP students. It aims to describe the studies briefly, contrast the contributions with each other and identify best practices from the literature. As such, data is presented in this manner, in Table 3.6.

Pedagogical Contribution	Study author(s) and reference number
Technical aspects of MT (Understanding the technical elements of machine translation systems allows translators to understand automatic metric scores and correlate them with human judgement)	Korošec (2011) [1], Xu and Wang (2011) [2], Doherty and Kenny (2014) [7], Rossi (2017) [12], Moorkens (2018) [16], Cid-Leal et al. (2019) [17] Bulut (2019) [19], Rothwell and Svoboda (2019) [20], Krüger (2021) [30], Krüger (2022) [32]
Critical/Reflective use of MT (Principles that make translators reflect about the output of MT and its evaluation procedures)	Venkatesan (2018) [13] Moorkens (2018) [16] Olkhovska and Frolova (2020) [24] Loock (2020) [25] González Pastor (2021) [28]

Table 3.6 continued from previous page

Pedagogical Contribution	Study author(s) and reference number
Translators as key stakeholders in evaluation (Centring translators in the process as key individuals for conducting evaluation)	Massey and Ehrensberger-Dow (2017) [10] Doherty et al. (2018) [14] Castilho et al. (2018) [15] Bulut (2019) [19]
Evaluation design (Devising comprehensive evaluation plans that account for biases, risk mitigation, etc.)	Schnell and Rodríguez (2017) [11] Doherty et al. (2018) [14] Guo (2020) [26]
Linguistic analysis Error detection (Different types of linguistic analysis guided by error typologies)	Venkatesan (2018) [13] Loock (2020) [25] González Pastor (2021) [28]

Table 3.6: Pedagogical contributions of studies under Rationale organised by authors and respective reference numbers

A focal point of several of these pedagogical studies, such as Korošec (2011), Xu and Wang (2011), Kenny and Doherty (2014), is the SMT paradigm, as well as the shift to NMT (see Rossi (2017) and Moorkens (2018)). The SMT studies emphasise that the technical aspects of SMT were paramount for educating translators about the systems and highlighting the human dimension, and the teaching of its technical aspects was critical for translators in training. As the field evolved, this understanding of the technical and linguistic understanding of MT systems broadened into what is now termed MT literacy[4] (Krüger, 2021). MT literacy may also involve a deeper understanding of how human judgement of MT quality correlates with AEMs in the context of MTQA education. Notably, the NMT paradigm posed new challenges, and so the ability to tailor evaluation for different contexts became

[4]Bowker and Ciro (2019)

indispensable, something that has been suggested since the SMT paradigm (Kenny and Doherty, 2014, p. 285).

NMT-focused studies, like those by Venkatesan (2018), underscore translators as central stakeholders, a sentiment also echoed in the training by González Pastor (2021). In this paradigm, translators, even during training, demonstrate a refined ability in error detection, a skill observable in both SMT and NMT teaching contexts (Rossi, 2017; Moorkens, 2018). This expertise becomes even more significant with NMT's unique error types (Breyel and Grass, 2020).

Loock (2020) aims to teach students to identify unique features produced by MT. Loock (2020, p. 165) does so by identifying Machine Translationese, which is the underuse or overuse of certain linguistic features due to algorithmic bias arising from training data or the algorithm itself, and shows that "by focusing on specific linguistic features, translators-to-be can be sensitised to the performance and limits of MT systems and therefore define their added value over the machine". This mirrors Moorkens (2018, p. 384) assertion that students can be "empowered to discuss relative merits and drawbacks" of MT. However, without this training, students might treat the output of MT uncritically, as tested by Olkhovska and Frolova (2020). Therefore, it is evident that the human expertise of translators can significantly enhance MT development when they act as evaluators. After all, translators have the ability to discern errors and inadequate MT output, such as when identifying linguistic features produced only by NMT (Loock, 2020) and preparing MT systems for the appropriate use (Cid-Leal et al., 2019) and can recognise the strengths, limits, weaknesses of MT with the value of expert translators (Bulut, 2019).

Castilho et al. (2018) discuss how translators have honed these skills both in industry settings and during their training. This positions translators as crucial stakeholders in evaluation, not just for their error detection prowess but also for their ability to contribute to other key aspects of HE, such as the qualitative feedback grounded on context, genre and purpose of text; nuanced capacity for interpreting contextual information (Bulut, 2019) that few AEMs are able to provide; and ethical

decision-making. Ethics can also play a crucial role in MT evaluation (Schnell and Rodríguez, 2017) as it aids the identification of risks, and it can be embedded in curriculum design (Massey and Ehrensberger-Dow, 2017).

Furthermore, fostering this synergy between humans and machines can cultivate a competence that equips translators to critically interact with automated systems (Bulut, 2019). Doherty et al. (2018) showcases and highlights the efforts in TS in embedding these principles through structured programmes, competences models, and extensive literature on MTQA. This is corroborated by Rothwell and Svoboda (2019) who acknowledge the increasing emphasis on quality assessment and metrics in training programmes. Doherty et al. (2018) further introduce the idea of selecting the most fitting evaluation based on various factors, such as workflow, text type, target audience, and perishability of texts - a skill that students can develop and report on during their training (Bulut, 2019).

In essence, discerning when and how to employ a comprehensive evaluation is a valuable skill for MT developers. Although "limited motivation to learn about and conduct TQA" (Doherty et al., 2018, p. 103) may represent a hurdle in development, introducing training to facilitate MTQA processes can promote the integration of quality assurance in different levels in workflows that are more sustainable in terms of time and costs, overcoming that particular difficulty. A practical recommendation is fostering research in MT development and evaluation by adopting interdisciplinarity (Guo, 2020), because "translation students and their teachers may need to acquire greater computing skills" (Kenny and Doherty, 2014, p. 289), demonstrating that learning from NLP has proven to be a recommendation for training in TS. As a pedagogical example, Krüger's studies (2021; 2022) equip translators with the skills to apply AEMs, while in terms of curricula, the European Master's in Translation (EMT) survey shows the effort of programmes to include MT evaluation through AEMs (Rothwell and Svoboda, 2019). Thus, a great interdisciplinary motivation can be placed upon doing the same towards NLP students, by equipping developers with HE skills, especially towards "the development of the human dimensions of

intuition, creativity, and ethical judgement" (Massey and Ehrensberger-Dow, 2017, p. 308) in relation to evaluation.

3.5.1.2 Resources (RE)

This subsection aims to identify the prevalent subthemes and topics of studies under the theme RE, addressing the question SRL-Q5 - What educational materials are used for MTQA teaching in TS and NLP education?. The studies in RE emphasise the tools, materials, and platforms used in teaching MTQA. While they might not delve deeply into pedagogical techniques, they offer insights into the tangible and intangible assets that are used in TS, which informs the selection and adaptation of materials that could be tailored for NLP students. These resources can also be used to show NLP students who wish to develop evaluation platforms what translators need. This section seeks to critically appraise the studies by contrasting the different resources presented, in order to pinpoint useful tools that can be potentially used to teach HE in NLP. It aims to describe the studies briefly, contrast the contributions with each other and identify best practices from the literature. As such, the data is presented in this manner in Table 3.7.

Pedagogical Contribution	Study author(s) and reference number
Free or Open Source Software (Recommendations in terms of resources or resources provided are free/open source)	Florez and Alcina (2011) [3] Post and Lopez (2014) [5] Krüger (2022) [33]
HE teaching (This resource focuses on methodologies for teaching HE)	Post and Lopez (2014) [5] Mitchell-Schuitevoerder (2020) [23]
AEM teaching (This resource focuses on methodologies for teaching the use of AEMs)	Doherty and Kenny (2014) [6] Krüger (2021) [30] Krüger (2022) [32]

Table 3.7 continued from previous page

Pedagogical Contribution	Study author(s) and reference number
HE + AEM teaching (This resource focuses on methodologies for teaching the correlation of HE and AEMs.)	Post and Lopez (2014) [5] Mitchell-Schuitevoerder (2020) [23]

Table 3.7: Pedagogical contributions of studies under Resources organised by authors and respective reference numbers

A few useful tools include an application created by Post and Lopez (2014) that integrates an evaluation-centric assignment module, where students rank translation outputs derived from a customisable dataset. While the default evaluation mode emphasises HE, it also provides an option to incorporate a reference translation for generating BLEU scores (Papineni et al., 2002). This design inherently promotes HE or a comparison between HE and AEM, making it an excellent tool for designing MTQA assignments, something that can inspire the development of MT evaluation programs. Krüger's studies (2021; 2022) use Jupyter notebooks to provide access to AEMs such as F-Measure (van Rijsbergen, 1979), BLEU, METEOR (Banerjee and Lavie, 2005), chrF (Popović, 2015), TER (Snover et al., 2006) and BERTScore (Zhang et al., 2019) with the goal of including NLP in translator training (Krüger, 2021, p. 11). The value of this approach can be demonstrated to MT developers through the careful study of the design of such applications that can be used to either introduce to or facilitate the use of AEMs by translators or other evaluators. This would allow NLP students to devise platforms that fit evaluation plans that blend HE with AEM scores that translators or other evaluators could use for an integrative evaluation.

Focusing on the use of HE in addition to AEM, Mitchell-Schuitevoerder (2020) presents a project-based approach using Dynamic Quality Framework (DQF) (TAUS, 2019) as the error typology model and an AEM to test an integration of MT within a CAT tool. The design of the activity and the brief of the assignment fosters eval-

uation approaches that make use of both HE and AEMs, which in the context of this thesis, is the kind of assignment that can be used to teach and emphasise HE as NLP students may possibly be more acquainted with AEMs.

Drawing connections between the rationale (Section 3.5.1.1) and resources for MTQA reveals a relationship between the principles and the practice. Open-source methodologies are important for MT research and education (Florez and Alcina, 2011). It resonates with Kenny and Doherty (2014)'s example of the importance of the Moses toolkit, which allowed primarily computer scientists to contribute with their own systems.

However, some software and programs can be difficult to use for translators to use, as their expertise in computing may be limited. It may lead to scenarios such as the one reported by Doherty and Kenny (2014) with the use of the Asiya-Online toolkit (González Bermúdez et al., 2012) which focuses on AEMs. According to Doherty and Kenny (2014), 20 of their 38 students abandoned the tool after attempting to use it, despite receiving the necessary training with the tool, and only after the toolkit had an improved interface were students able to use it. Thus, Krüger's (2021; 2022) solutions for translators can serve as an inspiration in terms of being easy to use for translators and for being open-source.

NLP students can benefit from Krüger's (Krüger, 2021, 2022) tools by learning to analyse what user interface (UI) aspects make them accessible to translators, in order to create platforms that facilitate evaluation. However, since the resources only account for AEMs, NLP students as developers must create or use a platform that also allows the use of both HE and AEMs. Rossi (2017) and Moorkens (2018) accentuate the inclusion of HE evaluation metrics and qualitative reflection in both SMT and NMT paradigms to understand the technical aspects of the different paradigms. Mitchell-Schuitevoerder's (2020) project-centric approach can be seen as a practical embodiment of the pedagogical suggestions emphasised by Doherty et al. (2018) of including evaluation types in workflows of different scopes.

3.5.1.3 Modality (MOD)

This subsection aims to identify the subthemes and topics of studies under the theme MOD, answering SRL-Q3 - How is MTQA taught in different teaching modalities in TS and NLP education?. These studies explore the different teaching modalities and tools to deliver content and instruction in teaching MTQA within TS. While not necessarily focusing deeply on pedagogical procedures, these studies explore the different modalities for delivering teaching MTQA within TS, that is, online versus offline, and shed light on the potential impact of each modality on student outcomes. This section aims to succinctly describe the studies, compare the modalities, and highlight their impact on TS student training, which, in turn, will feed into the methodology that is being developed in this thesis. It aims to describe the studies briefly, contrast the contributions with each other, and identify best practices from the literature. As such, data is presented in this manner in Table 3.8.

Pedagogical Contribution	Study author(s) and reference number
Hybrid training (This study provides insight on educational practices in the hybrid modality)	Rothwell and Svoboda (2019) [20] Rodríguez de Céspedes (2019) [21]
Online training (This study provides insight on educational practices in the online modality)	Xu and Wang (2011) [2] Rothwell and Svoboda (2019) [20] Rodríguez de Céspedes (2019) [21] Munkova et al. (2021) [34]
HE teaching (This study provides insight on how modality affects the teaching of HE)	Post and Lopez (2014) [5] Mitchell-Schuitevoerder (2020) [24]

Table 3.8 continued from previous page

Pedagogical Contribution	Study author(s) and reference number
AEM teaching (This study provides insight on how modality affects the teaching of AEMs)	Krüger (2021) [30] Krüger (2022) [32] Munkova et al. (2021) [33]
HE + AEM teaching (This study provides insight on how modality affects the teaching of HE correlated to AEMs)	Post and Lopez (2014) [5] Mitchell-Schuitevoerder (2020) [24]

Table 3.8: Pedagogical contributions of studies under Modality organised by authors and respective reference numbers

Designing an effective syllabus requires not only the careful selection of content but also considerations of the resources, values, beliefs, and the specific teaching conditions (Doherty and Kenny, 2014). Training on MTQA in TS also happens in hybrid training (Rothwell and Svoboda, 2019) and/or online training, and it can be a combination of asynchronous (recorded webinars and lessons or self-study via tools, assignments, wikis, etc.) or synchronous (live classes or exercises) elements. These modalities find applications both in academic settings (Rothwell and Svoboda, 2019) and the industry (Rodríguez de Céspedes, 2019).

Within the realm of the EMT programmes, Rothwell and Svoboda (2019) have observed a surge in the utilisation of online tools. These tools not only facilitate distance delivery of courses, but also promote collaborative projects, demonstrating the versatility and expansiveness of online education for MTQA. From a practical perspective, Munkova et al. (2021) developed an online teaching system for a MT course, incorporating evaluation elements focusing on AEMs (such as BLEU, WER and TER). Krüger (2021) emphasises that needs analysis is paramount, as the pedagogical practice of his study aimed to introduce NLP concepts such as word

embeddings and AEMs to students on an MA in Translation using Jupyter Notebooks, which can be used both asynchronously and synchronously, through a hybrid or distance format. It also shows that Krüger (2021) adapted the tools necessary for his students considering their potential lack of computing expertise.

The emergence of hybrid and online MTQA training studies (Post and Lopez, 2014; Munkova et al., 2021; Krüger, 2022) highlights the sentiments of Doherty and Kenny (2014) and Krüger (2021) that the foundation of any effective design and implementation of a syllabus requires understanding its context. This aligns with the research design in this thesis that aims to follow the use of ADDIE in order to understand the context of the NLP students and design a syllabus that will be relevant to them. Pedagogical research in TS demonstrates the adaptability of trainers in this evolving landscape of technology may affect the delivery of the teaching itself.

3.5.1.4 Teaching Procedures and Methods (TPM)

This subsection aims to identify the prevalent subthemes and topics of studies under the theme TPM, in order to partially respond to SRL-Q1 - How is MTQA taught in TS education? The pedagogical studies in the SRL were divided by their granularity, so the studies in this stage focus on the micro-level. In other words, it denotes studies that provide a close, detailed examination of individual teaching procedures, strategies and techniques, as opposed to broader investigations on curricula, which will be addressed in the next section. The purpose of this section is to capture elements that would be most relevant to teaching HE to NLP students with respect to teaching strategies. As such, data is presented in this manner, in Table 3.9.

Pedagogical Contribution	Study author(s) and reference number
HE teaching (Pedagogical procedures towards the teaching of HE)	Núñez (2019) [18] Bulut (2019) [19] Bogush et al. (2019) [22]

Table 3.9 continued from previous page

Pedagogical Contribution	Study author(s) and reference number
AEM teaching (Pedagogical procedures towards the teaching of AEM)	Munkova et al. (2021) [33]
HE + AEM teaching (Pedagogical procedures towards the teaching of HE correlated with AEMs)	Doherty and Kenny (2014) [6] Esqueda (2021) [30] Moorkens (2018) [16] Mitchell-Schuitevoerder (2020) [24] González Pastor (2021) [28]

Table 3.9: Pedagogical contributions of studies under Resources organised by authors and respective reference numbers

Several elements can be derived from the teaching approaches in TS to design training for NLP students. First, it is important to highlight the moments where humans participate in the evaluation. For instance, NLP students should be introduced to the concepts of adequacy and fluency as Doherty and Kenny (2014) and Moorkens (2018) did with their students. As well, they should run evaluations in different domains (Bogush et al., 2019), as students possibly have not been given training in the particular elements of text in different domains.

This is important because different domains have their own terminology, formatting and other elements. That helps testing the applicability of MT systems in different contexts. The aim of educating NLP students in evaluation scenarios in different domains, is to prompt NLP students to contemplate the subjective elements of the process as texts in different domains have varying linguistic conventions. NLP students may not be able to provide and are not expected to notice a nuanced observation like effected by expert evaluators in those domains. But it is important to establish that through classroom exercises coupled with examples from expert evaluators. This way, NLP students can be shown the importance of

the domain expertise of their evaluators. Secondly, for methodologies demanding deeper translation expertise, NLP students can learn about errors produced by MT and be tasked to identify issues such as contextual information missing (Bulut, 2019; Núñez, 2019) and inadequate or adequate language for their respective domain (Bogush et al., 2019; Bulut, 2019). Subsequently, in the context of this thesis, it would be expected that NLP students use their expertise with AEMs to capture how they may correlate or not with the HE approaches they have used.

Those tasks should integrate as many multiple evaluators as possible (Doherty and Kenny, 2014, p. 309) although it is acknowledged that there may be barriers, since it is not expected of NLP students to be multilingual like translators are. Structuring all those steps as a project aligns with the approach proposed by Mitchell-Schuitevoerder (2020), which provides an assignment design for group work to use HE and AEMs, and can further be broken down into task-based learning, adopted by Esqueda (2021), in which each task allows the sequencing of activities that can be adapted according to the student's needs and the LO they must reach.

Measuring students' perceptions of MT systems (Núñez, 2019; González Pastor, 2021), how confident they feel using MT systems and evaluation tools (Doherty and Kenny, 2014; Moorkens, 2018) are all pedagogical practices found in MTQA classrooms within TS. For NLP students, however, it is can be just as relevant to measure how confident they would feel to plan and implement evaluation for the MT systems they create. Particularly, diagnosing students at the beginning and at the end of the training (Núñez, 2019; Doherty and Kenny, 2014) complemented by assignments which can be used to elicit reflection from students (Moorkens, 2018; Esqueda, 2021) are approaches that will be used.

3.5.1.5 Curriculum and Syllabus (CSL)

This subsection aims to identify the prevalent subthemes and topics of studies under the theme CSL, addressing the SRL-Q1 - How is MTQA taught in TS?. Following the exploration of TPM studies, this subsection refers to the broad, overarching

perspective that encompasses entire courses, programmes, or curricular structures, focusing on a 'macro-level' perspective that shapes how MTQA is integrated and taught in TS. As such, data is presented in this manner in Table 3.10.

Pedagogical Contribution	Study author(s) and reference number
MTQA in a non-MT module (Consider the integration or discussion of MTQA in a non-MT module)	Korošec (2011) (2011) [1]
MTQA within MT module (Discuss/integrate MTQA within a MT module)	Dejica-Cartis (2012) [4] Doherty and Kenny (2014) [6] Schnell and Rodríguez (2017) [11] Rossi (2017) [12] Doherty et al. (2018) [14] Cid-Leal et al. (2019) [17] Rothwell and Svoboda (2019) [20] González Pastor (2021) [28]
MTQA as standalone training (This study discusses MTQA as its own training)	Farrell et al. (2017) [9] Rothwell and Svoboda (2019) [20] Rodríguez de Céspedes (2019) [21] Doherty et al. (2018) [14]

Table 3.10: Pedagogical contributions of studies under Curriculum and Syllabus organised by authors and respective reference numbers

It is established that training in MTQA happens both in formal teaching contexts as part of structured programmes and standalone courses and workshops (Doherty et al., 2018). The evaluation of MT in TS curricula has slowly been introduced as part of the use of MT (Korošec, 2011; Dejica-Cartis, 2012). MTQA practices explicitly emerged in curricula as an important aspect of MT use (Doherty and Kenny, 2014), considered as relevant content for critical approaches towards analysing MT output (Rossi, 2017), or as part of the workflow to build an engine (Farrell et al., 2017). Other studies advocate for the integration of MTQA as an important skill for

translators (Cid-Leal et al., 2019; González Pastor, 2021), focused on meeting industry needs (Schnell and Rodríguez, 2017). These studies suggest that MT/MTQA practices may not have to be reserved for a specific module within a programme, and should be integrated within other existing modules. Based on existing programmes in the EMT, it is demonstrated that MTQA appears as a standalone element in curricula (Rothwell and Svoboda, 2019).

Within the translation industry, there is no set curriculum, but rather, there is continuing professional development (Rodríguez de Céspedes, 2019), which fits the current needs of the company for the employee to learn, which happens to graduates from TS programmes (Schnell and Rodríguez, 2017). This approach to continuous development is echoed by González Pastor (2021), who stresses that translators have to deal with different phases, steps and agents with the further technologisation of the profession. However, that also means that more automation requires more attention towards human skills focused on personal and linguistic aspects, emphasised by Rodríguez de Céspedes (2019).

Furthermore, reproducible quality measures and clear guidelines from the industry are relevant to be included in MTQA teaching, as Doherty et al. (2018) recommend. Most importantly, the relevant aspect that can be applied for NLP curricula is to include the principle of being "aware of different types of evaluation to fit a particular project or workflow, and to consider whether a more innovative or agile approach could be used to replace another" (Doherty et al., 2018, p. 101), as MTQA should present objective, flexible (González Pastor, 2021), critical and reflective (Doherty and Kenny, 2014; Rossi, 2017) dimensions in different instances within the design of syllabi. As such, while NLP graduates are most likely to be acquainted with AEMs by their current programmes, equipping them with HE metrics and methodologies increases their choices to conduct comprehensive evaluation designs when necessary.

Thus, for a MTQA syllabus for NLP students, the following elements can be drawn from some modules in TS:

1. the inclusion of HE, such as adequacy, fluency, usability (Doherty and Kenny, 2014).

2. the inclusion of how HE correlates with AEMs (González Pastor, 2021).

3. the inclusion of detection of recurrent errors (Rossi, 2017).

4. human issues in MT, such as ethics and the role of the human in workflows (Doherty and Kenny, 2014; González Pastor, 2021).

While the degree of HE training will not be as extensive as it is for translators, it should provide NLP students with the skills to choose the most suitable design and procedures for MTQA. Mainly, after understanding the strengths, drawbacks, and the best integrations in workflows that HE metrics, methodologies and evaluator types (such as translators as expert evaluators) can provide.

3.5.2 Stage 2 - Natural Language Processing

This section focuses on synthesising the data of the SRL, focusing on the NLP field. The process of synthesis in NLP was limited to three studies [5] (see Table 3.2), all of them pertaining to the theme CSL. Considering there were no studies pertaining to RTL, RE, MOD and TPM, it is possible that studies of MTQA teaching are limited in NLP.

3.5.2.1 Curriculum and Syllabus (CSL)

This subsection aims to discuss how MTQA is present within studies themed as CSL in NLP, addressing SRL-Q2. How is MTQA taught in NLP education?. Contrary to the synthesis conducted in TS, the literature on NLP education primarily concentrates on the overarching curricular design, so there is no focus given to TPM, MOD, RTL or RE perspectives in this synthesis, as it can be seen in Table 3.11.

[5] Appendix B shows a full list of the studies which were associated with themes, with studies ordered chronologically and referenced by a number. The studies in each line were assigned to one or more themes depending on their contributions.

Pedagogical Contribution	Study author(s) and reference number
MTQA in Curricula	Martynova et al. (2018) [34]
MTQA in Short Courses	Alm et al. (2016) [33] Artemova et al. (2021) [35]

Table 3.11: Pedagogical contributions of studies under Curriculum and Syllabus organised by authors and respective reference numbers

Alm et al. (2016) designed a workshop for both computing and non-computing students, aiming to familiarise them with NLP applications. The workshop covered various aspects of NLP, including its history and applications such as information extraction and sentiment analysis. While MT was only one of the applications introduced, there was no specific focus on MTQA.

Artemova et al. (2021) developed a MOOC on NLP for Russian-speaking students, which included a dedicated section on MT. The MT section offered insights into MT systems history, advanced architectures and quality measures, with a particular outcome focused on human judgements. Specifically, the course touched upon MTQA introducing AEMs such as WER and BLEU. However, HE for MT was mentioned but not given many details or emphasis, indicating room for improvement in how the course covers HE.

Martynova et al. (2018) conduct a curricular analysis of three NLP programmes in Germany and four NLP programmes in the United States. From the analysis, it is shown that only three programmes offer MT as a module. In further investigation of the description of the modules of the programmes, there is no explicit mention that students are taught how to design or implement MTQA.

While institutions increasingly recognise the importance of NLP and its applications, the comprehensive teaching of MTQA is not prioritised. As evidenced by the courses designed by Alm et al. (2016) and Artemova et al. (2021), MT is often introduced as one application among many, with little to no focus on its quality assessment. This choice in curricula may risk producing systems that, while tech-

nologically and technically advanced, may not account for the quality expectations of real-world applications or diverse user bases in different scenarios, and may propagate errors and/or biases.

As the curricular analysis by Martynova et al. (2018) suggests, many NLP programmes may not emphasise quality assurance aspects of MT developments. It may lead to the development of NLP professionals who might be technically proficient but might lack the critical approaches required for a comprehensive approach towards MTQA, something that TS has been able to embed into the development of professional translators. TS as a field has been able to integrate MTQA proactively and holistically into its curriculum because of its focus on forging a tight relationship between the technological and the humanistic aspects, as the field itself recognises the indispensable nature of expert translators. Thus, TS is able to teach the critical use and evaluation of MT, NLP curricula should foster the critical development and evaluation of MT by reproducing successful practices that occur in TS.

3.6 Principles towards HE in NLP education

This section aims to establish the principles of HE in MTQA that can be used in NLP education based on the synthesis carried out throughout Section 3.5.2. Studies in TS have shown that a human-centred approach is capable of addressing progress in technology while still maintaining a critical tone (Kenny and Doherty, 2014; Rossi, 2017; Bulut, 2019; Olkhovska and Frolova, 2020; Loock, 2020), especially in estimating the appropriate scenario and domain through the evaluation process. Thus, I propose principles to apply in NLP education, followed by a brief discussion to demonstrate how the literature of TS will directly contribute to NLP education in the context of this thesis.

3.6.1 Principle 1: Emphasis on Human-Centric Evaluation in MTQA Training

A core tenet of MTQA found in TS is the irreplaceable value of humans in the interaction with MT, including in the evaluation process. From my understanding, NLP students, while technologically adept, benefit immensely from understanding HE from the point of view of translators. NLP students might not have the linguistic expertise of translators, but using HE approaches can enhance their comprehension of how systems can be evaluated. Specifically, being able to do so, prevents risks in the development of systems; risks that can lead to miscommunication, loss of reputation, or even injury or death, a typology discussed in-depth by Canfora and Ottmann (2018). In addition, NLP students can devise comprehensive MTQA plans and manage the use of evaluators in the most appropriate way, according to what each context may require, whether they are developing and evaluating a general MT system or a domain-specific one, as only humans provide nuance understanding of context. This principle also connects to TH, reflecting how philosophically I defend that humans have an indispensable role in MTQA, the SRL shows throught his principle that studies also advocate for the positioning of humans as the main feature in MTQA.

3.6.2 Principle 2: Adaptable and Contextualised MTQA Training

From a pedagogical point of view, training in MTQA requires flexibility and a needs analysis. The literature in TS demonstrates that groups have been taught under different circumstances and pedagogical interventions were adapted in the classroom and within the curriculum to address the students' needs and skills. In particular, it has been verified by TS teachers that it is beneficial to identify skills translation students may lack and integrate these skills into their modules or create separate training outside of the established curriculum through standalone courses. Moreover,

this type of training must be integrated from the perspective that the skills must be put in contexts where they are particularly useful and effective.

In the case of NLP students, teaching them how to conduct HE is important, but it must be contextualised that MTQA is an essential part of MT development and research. Thus, it must be demonstrated to identify the proper workflows and when and how to employ MTQA. Above all, the stages where HE and the evaluators (specifically translators) must be employed must be clear for the future-to-be MT developers.

3.6.3 Principle 3: Integrative Evaluation Platforms and Resources

NLP students acting as MT developers can connect technology and HE. Resources to carry out MTQA must be comprehensive, integrate HE with AEMs, and be user-friendly. Platforms and resources for evaluation that are intuitive to translators or other types of evaluators are part of comprehensive MTQA. Translators are experts in language, and may not have extensive technical expertise, thus, the way the evaluation is carried out must cater to this reality. For example, more recent platforms such as Machine Translation Evaluation Online (MATEO) (Vanroy et al., 2023) and MutNMT (Ramírez-Sánchez, 2023) aim to introduce non-experts to MTQA or provide accessible resources to MT developers and researchers and TS scholars.

3.6.4 Principle 4: Emphasis on Reflective Practice in the classroom

Methodologically and pedagogically, during their training, NLP students must be exposed to reflective tasks, as the nature of their work is already practical. Above all, it is beneficial for MTQA tasks to involve reflective reports, as NLP students need to think about their experience and explain how they implemented evaluation and why it was useful for the MT systems they are analysing, whether they are creating

one or evaluating an existing one. Additionally, in TS, the pedagogical interventions measure the students' capabilities through questionnaires and their perceptions of the content are measured to provide the teachers with an insight on whether the specific intervention was effective or not. In the context of the training fit for NLP students, two important aspects must be measured: whether they understand the importance of HE in MTQA and whether they demonstrate that they feel confident and capable of conducting HE in MTQA. The approaches used in TS consisted of questionnaires for perceptions, attitudes and self-efficacy in addition to the reflective reports, which will be adopted in the MTQA training designed for this thesis.

3.7 Curricular Analysis - MT in the NLP curriculum

As discussed in section 3.5.2.1, the SRL synthesis demonstrated that few publications in NLP reported how HE is taught or if it is taught at all. Due to these findings, it was necessary to further investigate the presence of MTQA in NLP programmes. As outlined in chapter 1, since the focus of the NLP training is on master's students, I believed it would be possible to find more information about how MTQA is designed for NLP audiences through the analysis of curricula. Curricula may present content, whether in broad terms or specific components (Seel and Dijkstra, 2004), that demonstrates the types of problems students would solve. They can also be perceived as the collection of official and unofficial resources resulting from the work of teachers and students (Woods et al., 2010). In the context of NLP education, Madureira (2021) states that students sometimes learn evaluation as they develop systems, but much of the content they acquire about evaluation is unstructured. With that in mind, a list of programmes from ACL Anthology was chosen. [6] The list of programmes is available in Appendix C.

[6] Available on: `https://aclweb.org/aclwiki/List_of_NLP/CL_courses`

3.7.1 Inclusion and Exclusion Criteria

Given the diversity in programmes and different designations across countries, I make no terminological distinction between some areas such as Computational Linguistics (CL) and NLP, for example. While there are studies that may claim the distinction of both fields (Tsujii, 2011), discipline boundaries are not easily defined (Luz, 2022), for instance, Allen et al. (2021) analyses curricula in AI encompassing machine learning and deep learning, which can be a particular interpretation. Similarly, using my own interpretation without emphasising on limiting fields and their boundaries, this curricular analysis aims to explore related or connected fields as much as possible, but still focusing on Computer Science as a field.

Evaluating curriculum requires defining the purpose of the evaluation and the criteria included (Brady, 2013), it can support decision making to innovate, design, develop, implement or even change its different aspects (Klenowski, 2010). As previously mentioned, the main goal is to identify the presence of MTQA in NLP curricula. In order to gather the most information out of the available resources, I devised different inclusion criteria, and included curricula are placed upon one of two categories: i) possibly has MTQA ; ii) explicitly has MTQA. The following are the inclusion criteria for **"possibly has MTQA"**:

- The syllabi available contained MT as content
- The syllabi available contained MT as a task under the umbrella of NLP

While the following are the inclusion criteria for **"explicitly has MTQA"**:

- The syllabi available contained MT Evaluation as content
- The syllabi available contained 'evaluation' as a learning outcome of the module
- The syllabi available contained a HE metric as content
- The syllabi available contained an AEM as content

Likewise, exclusion criteria were developed for **"not considering the module to have MTQA"**:

- The syllabi available do not mention MT or MTQA or evaluation at all
- The syllabi available may offer content that may be related to MT, such as Deep Learning, Neural Networks, Sequence-to-Sequence Models, Embeddings, Machine Learning, but has no obvious connection to MT evaluation
- The syllabi are not publicly available (e.g. only available through institutional access or through an LMS)

3.7.2 Results

The analysis consisted of choosing master's programmes from the list that could possibly have MT or MTQA as a module, whether they were compulsory or optional. The list was organised by name, the master's programme name, the country, the name of the module, and a field to write whether MTQA is a component of a module or not.

3.7.2.1 Programmes

In total, 56 programmes were analysed (list of universities can be found in Appendix C). The following programmes were analysed: Computer Science, Artificial Intelligence, Computational Linguistics, Informatics, Language Technology, Data Science and Artificial Intelligence, Cognitive Systems, Natural Language Processing, Digital Text Analysis, Informatics Engineering, Information and Data Engineering, Artificial Intelligence and Automation Engineering, Human Language Science and Technology and Informatics. Of the 56 programmes, 39 had components or descriptors that indicated the teaching of MT, which may include evaluation. On further investigation, 30 of these programmes made an explicit mention of MT or MTQA, requiring a deeper investigation. Based on that, I calculated that a proportion of

69.64% of the programmes possibly contained MTQA, and 53.57% of programmes explicitly contained MTQA.

3.7.2.2 Modules

After analysing the 56 programmes, several modules were analysed, with the times they appeared on programmes between parentheses:

- Natural Language Processing (18 times)
- Machine Translation (5 times)
- Machine Learning (2 times)
- Advanced Natural Language Processing (2 times)
- Themes in NLP and Language Technology
- Applied Artificial Intelligence
- Introduction to Machine Learning for Language Applications
- Natural Language Understanding
- Language Processing Technologies
- Transformers
- Applied Machine Learning
- Deep learning with neural networks
- Statistical Machine Translation
- NLP Applications
- Competing in Machine Translation
- Basics in Machine Translation
- Advanced Machine Translation

- Current Topics in Machine Learning
- Computational Linguistics in Depth

When analysing these modules, it was noted that all modules named 'Natural Language Processing' and 'Machine Translation' either contained evaluation as a LO or offered evaluation as content in the publicly available syllabi. Under the 'Natural Language Processing' modules, MTQA was present with either MT described as a task and/or describing 'evaluation of tasks' as a LO. While for 'Machine Translation' modules, the descriptors were more specific, mentioning that several aspects of MT would be taught, from the training of systems to their evaluation. Some modules covered specific evaluation approaches, including AEMs and HE metrics, which can be seen in the next subsection.

3.7.2.3 Metrics

An examination was conducted on the programmes that featured specific metrics as content within the analysed syllabi. Of the 29 programmes that explicitly referenced MTQA, there were four mentions of AEMs, and three mentions of HE.

The following AEMs were found in the curricula:

- BLEU
- chrF
- chrF++
- BLEURT
- BERTScore
- LLM-based metrics
- Reference-based AEMs (Matching and Edit Distance)
- Referenceless AEMs

- COMET
- BERT

Four programmes out of the 39 programmes that have evaluation contain AEMs. Thus, proportionally, there is 10% explicit presence of AEMs across the programmes. BLEU, BLEURT, and chrF are the most commonly cited AEMS within the curriculum. One programme chose to categorise metrics as either Reference-based or Referenceless, which could aid curricular design by not committing to a specific metric, reflecting the continuous evolution of the field. It is noteworthy that AEMs were exclusively noted in modules entitled "Machine Translation", and were not specifically included in any "Natural Language Processing" modules. This is likely because, in general NLP modules, other AEMs are taught for various tasks, and thus there is no explicit focus on metrics commonly used for MT.

The following HE metrics were explicitly found in the curriculum:

- Adequacy/Fluency
- Comprehensibility
- Ranking
- Scoring
- Marking

Two programmes out of the 39 programmes that have evaluation contain HE metrics. Thus, proportionally, 5% of all the programmes explicitly incorporated HE. Adequacy and fluency are the most frequently encountered HE metrics explicitly mentioned within the curriculum and feature in all modules titled "Machine Translation". Other evaluation measures, such as IAA, are included in a module titled "Natural Language Processing", which covers a range of NLP tasks beyond MT. It is worth mentioning that the taxonomy of errors varied between programmes. This probably occurs because the taxonomy in the literature also varies, as examined by Huidrom and Belz (2022).

3.7.3 Discussion

Based on these results of the curricular analysis and the results of the SRL, several inferences can be drawn. Evaluation is likely to be included in general NLP modules; however, there may not be a pronounced focus on MTQA due to the extensive range of NLP tasks that NLP educators must address to cover fundamental aspects of development. Some programmes devised solutions by creating dedicated modules to focus solely on evaluation, with titles such as "Natural Language Processing (Extended)" or "Advanced Natural Language Processing", allowing educators to devote more attention to various facets of NLP that previously received less emphasis.

Further investigation of other modules reveals that while MT modules typically include evaluation, there are modules with different primary objectives that also incorporate evaluation. For instance, the "Competing in Machine Translation" module prepares students to develop and assess MT systems for competitions such as the Workshop on Machine Translation (WMT), an annual event that draws MT researchers and companies to develop and evaluate MT systems.

From this analysis, several principles emerge that may present challenges for NLP educators, based solely on the curriculum:

- When NLP educators wish to place greater emphasis on evaluation (or MTQA), they may choose to offer optional modules specifically focused on MT or dedicated NLP evaluation modules.

- Programmes often have distinct focuses; some prioritise MT as a task, while others may favour different NLP applications. For instance, several "Natural Language Processing" modules do not highlight MT as a task among others, suggesting that more time might be allocated to other areas of NLP.

- Modules that concentrate on other aspects of AI, such as Deep Learning or Transformers, likely place a greater emphasis on the technical dimensions of system development.

It is important to acknowledge that this curricular analysis has several limitations. For instance, some syllabi were not publicly accessible, requiring institutional access for detailed curricular information. At other times, programmes lacked detailed content descriptions in their modules, making it challenging to determine whether they included MTQA or not. Conversely, other programmes made slides publicly available, facilitating a closer examination of the content that confirmed the inclusion of MTQA.

To verify or further investigate beyond this analysis and its limitations, please see Phase 1c in Section 4.1. Engaging directly with educators allows for the collection of insights that were unattainable in this phase. Although this phase has augmented the results of the SRL, highlighting the scarcity of dedicated curricular space for MTQA within general NLP modules, it reveals the need for more structured educational content on MTQA education for NLP audiences.

3.8 Conclusion

This SRL attempted to respond to several SRL-Qs in order to address Research Question 2a: "What content related to MTQA is currently taught in postgraduate NLP and Translation Studies programmes, if any?". The first SRL-Q, "How is MTQA taught in TS education?", was answered through a synthesis of studies themed on different aspects of curricula, syllabi and classroom practices. The analysis of TPM themes highlighted the importance of teaching the deeper understanding of both technical and linguistic aspects of MT systems for the purpose of evaluating them, such as the combination of correlating human judgement with AEMs with the expertise of translators, to achieve the advanced detection of errors and the critical analysis of MT output to diagnose its performance. That can be observed in the structured discussion in the TS studies themed with CSL, showing how MTQA has been integrated in the curriculum of programmes or how it can be an important aspect of modules currently taught in TS programmes.

The second SRL-Q, "How is MTQA taught in NLP education?" was explored

through a short review of two educational interventions (Section 3.5.2.1) and a curricular analysis (Section 3.7). The findings suggest a notable gap in how HE in MTQA is addressed within NLP education. While certain initiatives demonstrate how MT is introduced as part of NLP education, they do not emphasise a human-centric evaluation of MTQA and still leave many questions about how to teach MTQA to NLP students.

The third SRL-Q, "How is MTQA taught in different teaching modalities in TS and NLP education?", was explored through an examination of different teaching approaches including hybrid and online training in TS (Section3.5.1.3). The literature indicated a flexible and adaptive approach to MTQA education tailored to the specific conditions, resources, and the evolving technological landscape. Assertions of scholars showed how TS has been able to effect well-rounded and context-driven interventions in MTQA teaching in both modalities, although not much focus has been given in-depth to the aspect of modality in education.

The fourth SRL-Q, "What are the rationales for MTQA in TS and NLP education?", was examined through various studies that captured the transition from SMT to NMT (Section3.5.1.1). The findings highlight the need to provide a technical understanding of MT systems to better develop and perform a human-centric evaluation of the output. This focus in instruction ensures that translators become adept at error detection and fluency and adequacy assessments. Other considerations in some studies involve embedding ethical elements and interdisciplinary approaches demonstrating that translators benefit from computational skills such as the ability to use AEMs and interpret their results.

Finally, the fifth SRL-Q, "What educational materials are used for MTQA teaching in TS and NLP education?", was addressed through the analysis of some resources that emphasised HE, AEM and a combination of both (Section 3.5.1.2). While the tools analysed were aimed primarily at enhancing translators' evaluation skills, they serve as examples of tools that NLP students can develop. By studying the design and functionality of these tools, NLP students can incorporate intuitive

evaluation platforms that encapsulate both HE and AEMs.

In concluding this analysis, it is essential to address the inherent limitations due to the limited number of studies available in NLP teaching, specifically in the teaching of MTQA. The scarcity of research on NLP curricula with a focus on MTQA might signal two potential scenarios. Firstly, it may be indicative of an area that has not been extensively explored. The rapid technological advancements in the NLP field justify greater attention towards pedagogy. For example, Doherty et al. (2018) mention that TS scholars and teachers must remain up to date or even ahead of swift technological developments. Consequently, the importance of effective and contemporary pedagogical interventions, with an emphasis on quality assurance and risk prevention measures like MTQA becomes even more pronounced. While this chapter provides observations from the available data, the actual depth of MTQA instruction in NLP curricula may be underrepresented. The second scenario could relate to a possible reporting gap, with institutions or researchers not adequately documenting or publishing their curricular innovations in this area. This leads to the need for a more careful and guided curricular analysis, as carried out by Martynova et al. (2018) and later performed in Section 3.7.2, to gain more insight into how content is reflected in curricula.

In the context of this SRL, the curricular analysis performed in section 3.7.2 complemented the information retrieved from the data retrieval and synthesis. It identified the prevalence of explicit MTQA evaluation instruction when MT is the primary focus or emphasis. Additionally, it highlighted the potential curricular challenges associated with teaching evaluation across all NLP tasks, not solely MT. The analysis revealed a potential and significant disparity in the way evaluation is integrated into NLP curricula. While MT-focused courses often incorporate detailed assessment methodologies focused on depth, broader NLP courses may lack comprehensive evaluation components, which could hinder students' understanding of quality assurance in MT. This discrepancy underscores the need for the structuring of evaluation across NLP-related modules, at least in relation to MT.

Chapter 4

Methods - "Design"

This chapter will describe the methodology for Phase 1c (Survey with Teachers - Section 4.1), Phase 2 (Design Blueprint - Section 4.2), Phase 4 (Syllabus Implementation - Section 4.3, and Phase 5 5 (Evaluation - Section 4.4), as can be seen in Figure 4.1.

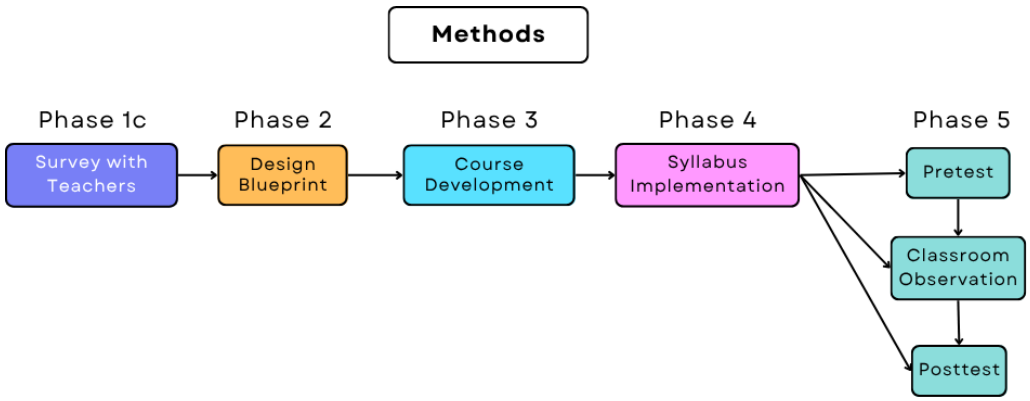


Figure 4.1: Figure showing the order of the methodology followed for each of the remaining phases

These phases are essential for the implementation and evaluation of the training on MTQA. Phase 1c impacts the subject matter expertise feeding into the design of the course in phase 2, so it can be developed in phase 3, implemented in phase 4 and evaluated in phase 5. Phase 1c findings describe a survey with teachers, which identifies the content NLP students should interact with in the training. As the last stage of the needs analysis, it actively impacts the core design and development aspects of the training. Phase 2 describes what elements of design will be used to

develop the training, in addition to providing the structure of the methods for phase 3, 4 and 5. Note that Phase 3, due to its scope and granularity, will be described fully in Chapter 5. All stages of data collection in this chapter were submitted to the Humanities & Social Sciences Faculty Research Ethics Committee under the reference DCU-FHSS-2023-015 and were approved as a low-risk project.

4.1 Phase 1c - Survey With Teachers

This section will describe the design of the phase that aimed to ask lecturers from TS and NLP about MTQA education. In participant-oriented research, participants can serve as agents involved in the research process itself (Saldanha and O'Brien, 2014). Here, trainers are participants. Needs analysis precedes the design of instruction. To identify the content that should be included in the training, subject matter experts (SME) must be consulted (Dick et al., 2021).

SMEs, lecturers in TS and NLP, are key stakeholders and their extensive experience in teaching the topic of MTQA allows suggestions and constructive feedback. SMEs are useful to provide information on what content should be taught, the sequence of contents, and the ideal conditions for teaching (Brown and Green, 2024, p. 68-69). To reach the SMEs, I recruited them via an online questionnaire distributed on social media and by email. According to Saldanha and O'Brien (2014, p. 151), surveys refer to the design of the study, while questionnaires are research instruments. The advantage of an online survey includes reaching many participants and being quick and cost-effective (Cohen et al., 2018, p. 361-362), which can be essential when the ID professional is under time constraints to develop a curriculum (Mattoon, 2005). However, on the downside, low responses can be an issue (Creswell, 2015). Foreseeing this issue, I designed a questionnaire with open-ended questions to gather detailed insights from the participants who responded to the questionnaire. Even with low numbers, in-depth qualitative insights can be useful.

With this in mind, the following are the two aims for the survey:

- Identify educators' attitudes towards MTQA.
- Identify what approaches and content can be taught to foster HE in MTQA.

4.1.1 Survey Setup

4.1.1.1 Participant Profile

The analysis of curricula showed that MTQA was taught to a diverse range of students in different programmes and by lecturers from fields beyond TS and NLP, including CL, AI, and human-computer interaction (HCI). Although terms are sometimes used interchangeably (Kühl et al., 2019), they have their own meanings among different communities. Therefore, this was considered when designing the questions and when recruiting the participants.

Participants were recruited through three different means:

i) Social media: Via a public post shared on Linkedin.

ii) Email collected from curricular analysis (phase 1b) of lecturers teaching MTQA or evaluation in NLP.

iii) Email collected during SRL from publications by scholars who published research on MTQA teaching.

The group selected for this study is a convenience sample as the study is small-scale and not randomised, as educational studies can be non-probabilistic in nature (Cohen et al., 2018), and the purpose of the investigation is not to generalise the findings but to use the convenience and availability of the participants (Creswell and Clark, 2017).

4.1.1.2 The Research Instrument - Questionnaire

Following *RQ3. "What can be drawn from TS when teaching MTQA for NLP postgraduate audiences"*, the questionnaire aimed to find adequate subject matter content and ideal teaching conditions. The results of Phases 1a and 1b provided me

with a direction of what types of questions should be asked. For example, I wanted to understand what specific content needed to be included, because curricula publicly available online did not describe details: I wanted to investigate whether HE and AEMs are taught together, and what specific metrics are taught. I also wanted useful contextual suggestions about the course type, the duration, and the teaching methodology of the classes.

The questionnaire (available in Appendix D) consisted of closed and open-ended questions organised in categories based on Creswell (2015):

- personal questions, such as country, level of education, field (Q1, Q2, Q3, Q4, Q5, Q10, Q11, Q12, Q13, Q22)
- attitudinal questions, such as the importance of incorporating HE in MTQA teaching, trends in evaluation (Q8, Q9, Q14, Q15, Q16, Q17, Q23, Q24, Q25, Q26, Q27, Q28)
- behavioural questions such as hours per week spent on teaching or whether the participant teaches quality assessment of other NLP tasks (Q6, Q7, Q18, Q19, Q20, Q29, Q30, Q31, Q32, Q33, Q34)

Closed questions offer the participant the practicality of knowing the range of responses they can choose, in addition to being useful for statistical testing, while open questions offer more depth by exploring attitudes and behaviours further. The questionnaire was hosted on Qualtrics[1], the survey platform licenced for university students and staff. Qualtrics is recommended for questionnaires in a survey study for its user-friendliness and presentation (Fielding et al., 2016, p. 60-61).It also allows survey settings that are recommended in the literature, such as adding a progress bar, testing the presentation of the survey on different devices (desktops, tablets, phones), and a screen thanking the participants once the survey is over, among other suggestions (Cohen et al., 2018, p. 365-366). Using the Qualtrics Survey flow functionality, I designed the questionnaire to have five blocks, as Figure 4.2 shows:

[1]Information retrieved from https://www.dcu.ie/iss/qualtrics. Accessed on 15/11/2024

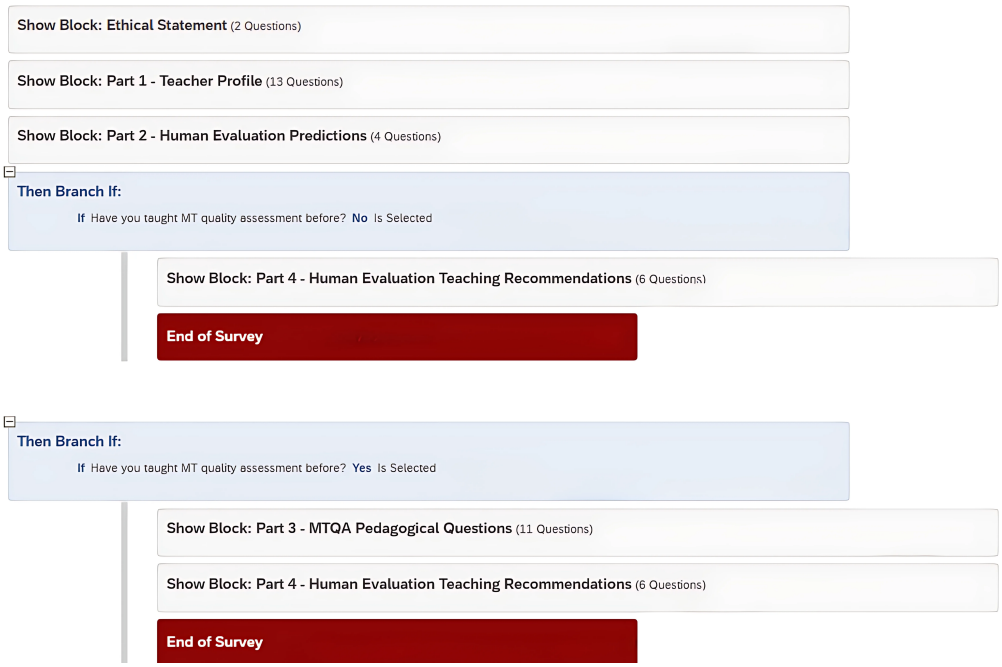


Figure 4.2: Flowchart showing the five blocks, Ethical Statement (Block 0), Part 1 - Teacher Profile (Block 1), Part 2 - Human Evaluation Predictions (Block 2), MTQA Pedagogical Questions (Block 3), and Human Evaluation Teaching Recommendations (Block 4)

This flowchart extracted from the platform shows the branching paths according to the participants. The first block contains the ethical statement that is necessary before any data is collected, hence why it is considered Block 0. After the participant accepts the terms in Block 0, they are allowed to proceed to the next block. Block 1 aims to identify the profile of the participants and pedagogical aspects of their work, such as workload and type of student they teach (such as undergraduate and postgraduate). Block 2 has questions related to the educator's attitudes towards HE and offers a branching path. If participants answer the question shown in the figure with "yes", they are led to Block 3 and 4. If not, they only respond to Block 4. Block 3 has more in-depth questions related to MTQA pedagogy, while Block 4 aims to collect suggestions on content to be included in the course and pedagogical elements such as type, modality, and duration of the course.

4.1.2 Data Analysis

4.1.2.1 Block 1 - Teacher Profile

The survey collected responses from 28 educators across different fields[2]. Q1 and Q2 were questions about contact details, which became pseudonymised data. Q3 asked the country where the educators taught, with the following responses, as image 4.3 shows:



Figure 4.3: Graph showing the countries of the participants

For **Q4 - Fields of work** participants could choose multiple fields. 18 participants chose TS as their field, of the 18, two also chose Speech Processing, one participant chose HCI, and another chose Computer Science in the field "Other". Five participants chose CL and five participants chose NLP, as seen in graph 4.4.

[2]In this Survey, no data is statistically significant, as the statistical power is low with N = 28.

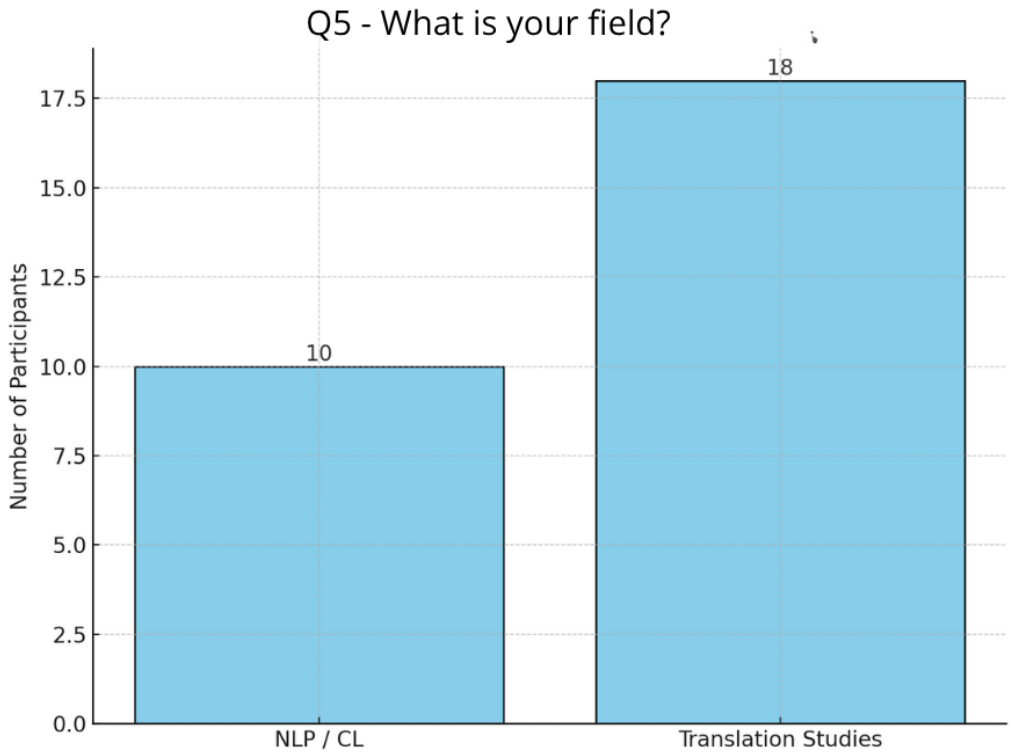


Figure 4.4: Graph showing the field of the participants

The distinction of CL and NLP was not made, because boundaries may not be easily defined between the two terms (Luz, 2022), leading to the grouping of responses from CL and NLP teachers as one single group. In fact, one of the participants responded in the "Other" field with "Why split NLP and CL?", reinforcing the blurriness of this distinction, leading to the grouping of their responses.

In Q5, when asked about their level of education, 25 educators responded with PhD, and 3 educators held a master's degree. In Q6, responding to number of hours spent teaching per week, Table 4.1 shows the average:

Field	Minimum	Maximum	Mean	Std Deviation	Variance
NLP	4.00	20.00	7.30	4.61	21.21
TS	2.00	12.00	6.35	3.25	10.58

Table 4.1: Descriptive statistics for hours spent on teaching by field

In Q7, when asked about their main work activities other than teaching, graph 4.5, describes their tasks:

Across all participants, the majority of educators' time is spent on research, accounting for 60.0% of primary work activities. This is followed by administrative

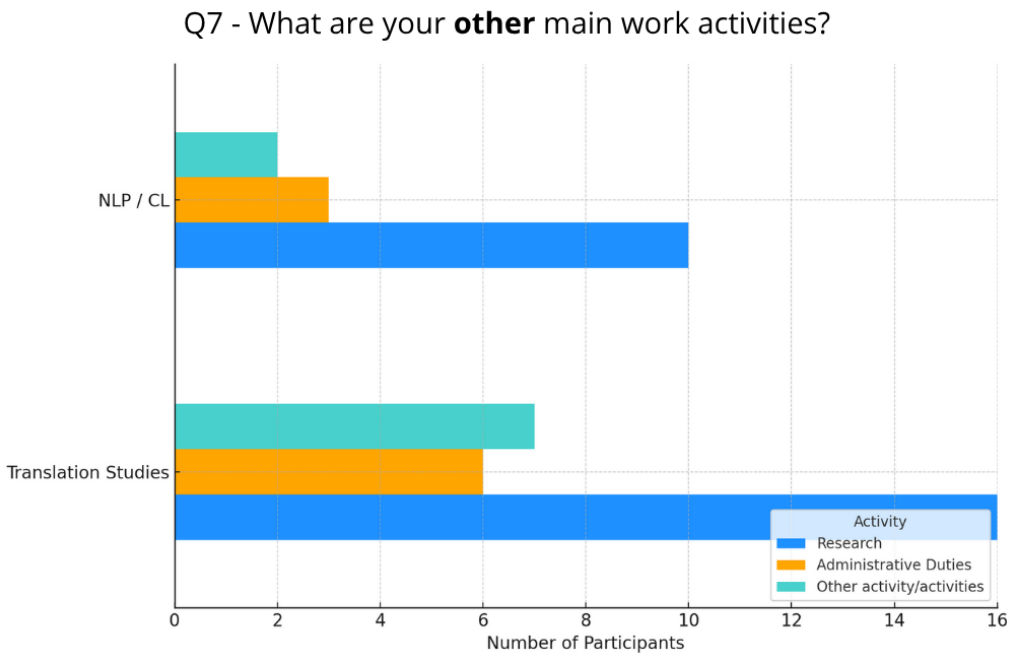


Figure 4.5: Graph showing the activities other than teaching between participants, including research, administrative duties and other activities

duties (20.0%) and other diverse activities (20.0%), which include tasks such as translating, editing MT output (for pedagogical purposes or freelance work), consulting, and organising conferences.

Analysing the data by field, educators in NLP/CL reported spending a greater proportion of their time on research (66.7%), with administrative duties comprising 20.0% and other activities 13.3%. In contrast, those in TS spent 56.7% of their time on research, the same proportion (20.0%) on administrative work, and a higher share (23.3%) on other professional tasks.

When asked about their influence over the curriculum (Q8), educators could choose from the lowest amount of influence ("No influence") to the highest ("Comprehensive Influence"), which can be seen in graph 4.6.

The distribution of influence over the curriculum is broadly similar across educators in CL/NLP and TS. In both groups, a majority report having at least some level of influence, with comparable proportions indicating significant or comprehensive involvement in shaping syllabi and teaching methods. This suggests that, regardless of disciplinary background, most educators perceive themselves as having moderate

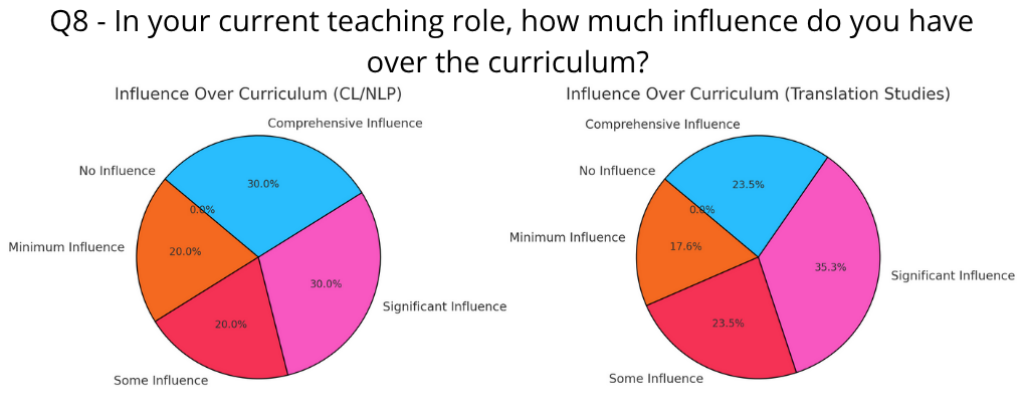


Figure 4.6: Distribution of educators' reported influence over the curriculum, grouped by academic background: CL/NLP and TS. The levels of influence are ordered from lowest to highest: No Influence, Minimum Influence, Some Influence, Significant Influence, and Comprehensive Influence.

to high agency in curriculum development. When asked for further comments (Q9), educators mentioned:

- Syllabi are officially approved by the faculty and not changed by the lecturer themselves, with some influence over the design but not the final approval, mentioned by one participant.
- Syllabi can be modified in terms of content and teaching methods, with the addition of including guest speakers as an autonomous decision, mentioned by 9 participants.
- Syllabi and entire programmes can be proposed and modified, but administrative, financial and regulatory constraints must be approved by the university at levels of faculty, university and ministry, mentioned by one participant.
- One educator reported that they hold their own localisation school, which allows full freedom, mentioned by one participant.

Responses to Q10, concerning the type of students taught by educators, can be seen in figure 4.7.

Both CL/NLP and TS educators teach at multiple academic levels, with undergraduate and postgraduate being the most common. CL/NLP educators are slightly

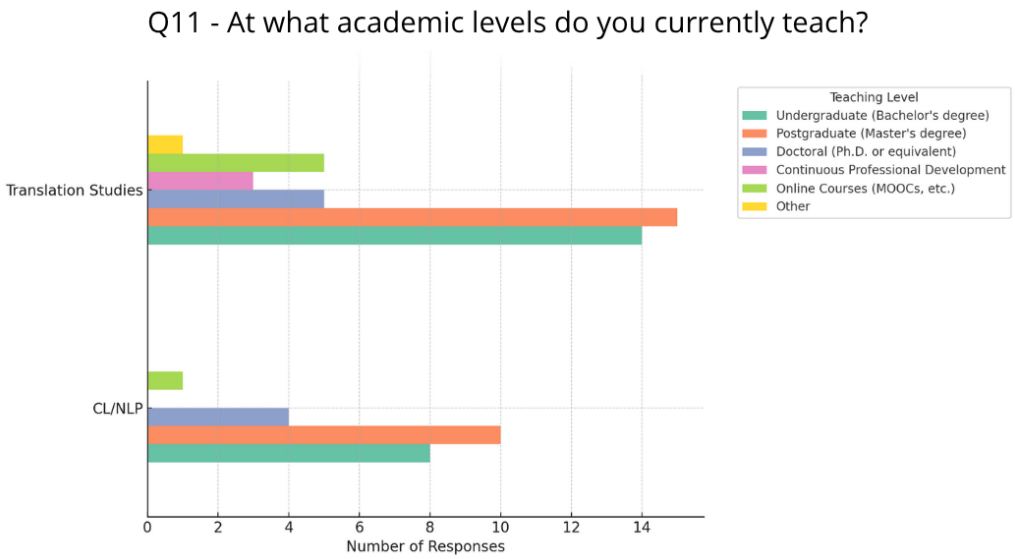


Figure 4.7: Bar chart showing the number of educators teaching at different academic levels, grouped by CL/NLP and TS. Each bar represents the frequency of responses for a given level, including Undergraduate, Postgraduate, Doctoral, Continuous Professional Development, Online Courses, and Other. Respondents could select multiple levels.

more concentrated at higher academic levels, while TS educators show broader involvement, including more in professional development. This reflects subtle differences in how each group is positioned within academic and training contexts. Q12, then, was concerned with the modality of the training they deliver, as Figure 4.8 shows.

CL/NLP and TS educators both show a strong reliance on in-person teaching, suggesting that face-to-face delivery remains the dominant mode in both fields. However, TS educators appear more engaged with flexible teaching modalities, reporting higher use of both hybrid and online modalities. This may reflect broader access needs or pedagogical traditions in TS, whereas CL/NLP educators remain more concentrated in conventional delivery, possibly tied to institutional structures, technical course requirements or number of students.

In Q13, when asked whether educators have taught MTQA before, 21 participants chose “Yes” and 7 chose “No”, broken down by field in Figure 4.9.

After Q13, Block 2 starts.

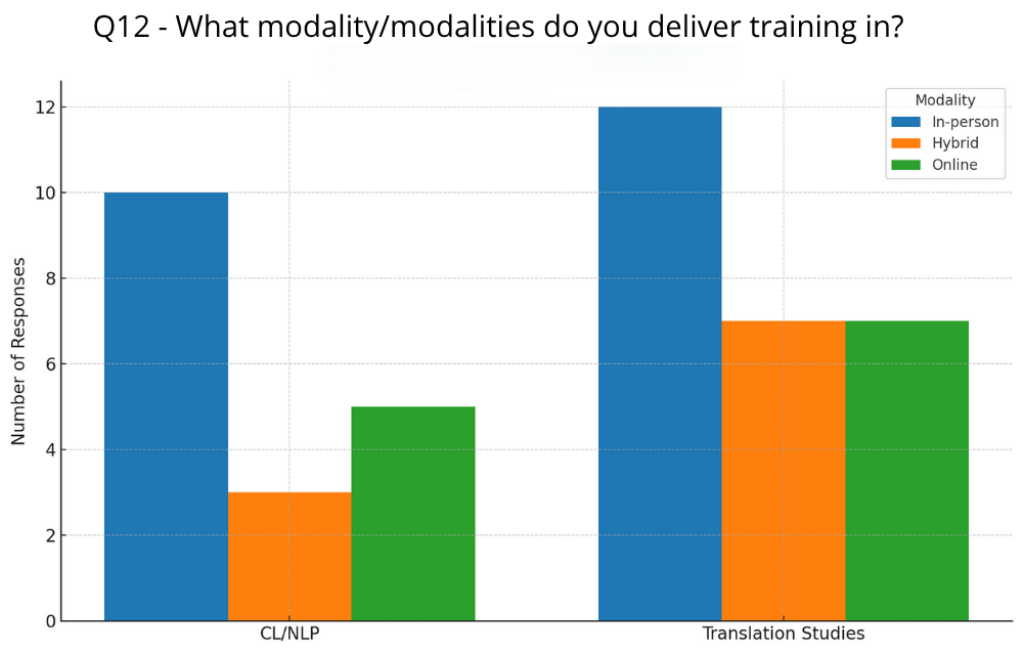


Figure 4.8: Teaching modalities reported by educators in CL/NLP and TS. The chart shows the number of respondents teaching in each format: In-person, Hybrid (a mix of online and in-person), and Online

4.1.2.2 Block 2 - Human Evaluation Predictions

In Q14, participants were asked to rate the significance of incorporating HE in the development of MT systems, as seen in Figure 4.10.

Most of the respondents in both CL/NLP and TS consider teaching MTQA to be highly important. In the CL/NLP group, responses are concentrated at the upper end of the scale, with seven out of ten rating it as "Extremely Important" and two as "Very Important." Only one respondent selected "Moderately Important." Similarly, the TS group demonstrates a strong consensus, with twelve out of fourteen rating it as "Extremely Important" and the remaining two selecting "Very Important." No participants in either group rated the topic as unimportant. This distribution indicates a shared recognition of the relevance of MTQA in both disciplinary contexts.

Q16 asked educators about concepts or topics that would become trends in HE of MT, in Figure 4.11.

The data show that both CL/NLP and TS educators foresee a range of trends,

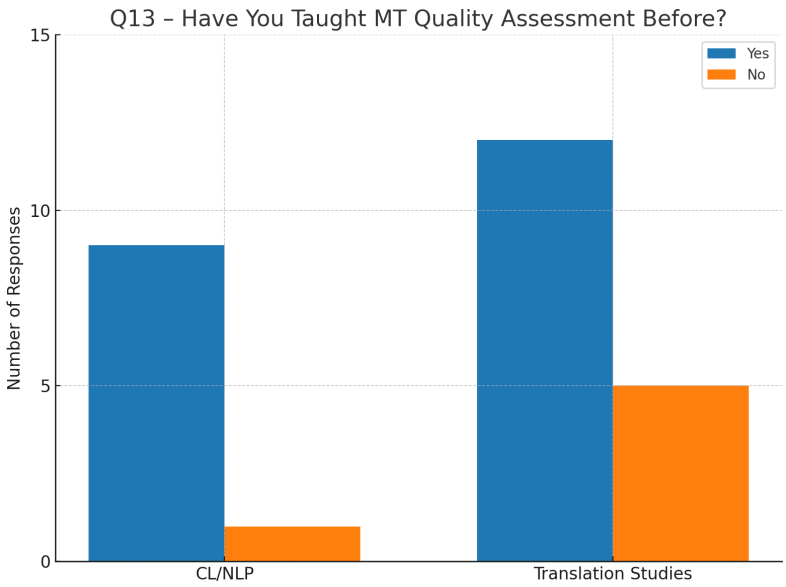


Figure 4.9: Number of respondents who have taught MT quality assessment, grouped by disciplinary background (CL/NLP and TS).

with overlap across some categories. Context-aware evaluation, UX[3], and ethics are among the most frequently selected trends in both groups, suggesting broad agreement on emerging priorities. TS respondents more frequently selected options across nearly all categories, particularly for context and multimodality, indicating slightly broader engagement with these evolving areas. Crowdsourcing was notably absent among CL/NLP responses, suggesting reservations about relying on this type of evaluation, considering potential quality control issues (Knowles and Lo, 2024).

Q17 asked educators what constitutes a comprehensive evaluation of an MT system, which was responded to differently by each group. 10 participants in the NLP group responded:

- Taking into account the data used (quantity and quality), size of the model (number of parameters), memory footprint, generalisation power, and robustness to domain.

- Evaluation with the intended purpose of the MT system.

[3]The survey presented no definitions of the concept of UX to participants, so the understanding of the participants may have varied. In this thesis, I follow the definition of UX provided by International Organization for Standardization (2019) where it is related to the perception or response of an end user of a product, system, or service.

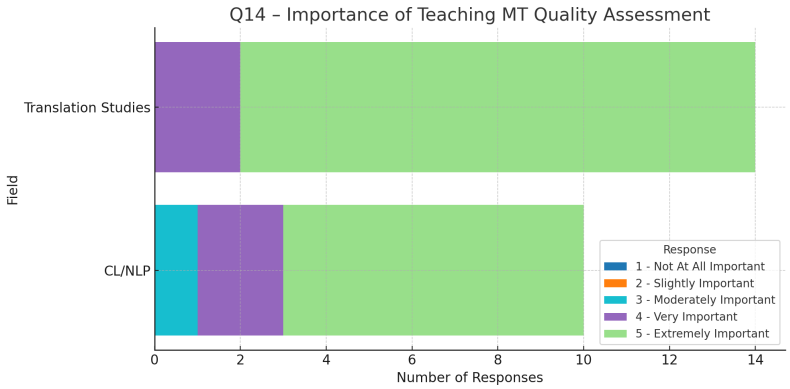


Figure 4.10: Perceived importance of teaching MT quality assessment, grouped by disciplinary background (CL/NLP and Translation Studies). Responses were collected on a 5-point Likert scale: 1 – Not at all important, 2 – Slightly important, 3 – Moderately important, 4 – Very important, and 5 – Extremely important.

- A combination of HE and AEMs, with task-based evaluation.
- A combination of domain, quality fit for purpose, and UX.
- A combination of specific metrics, such as adequacy, fluency, context-level aspects, and error analysis.

14 participants in the TS group responded:

- Specific language features, such as genre, style, terminology, and clarity.
- Specific HE metrics, such as adequacy, fluency, error typology, usability.
- Transparency with data, integrability with workflows.
- The use of the rationale of the Skopos theory for the intended purpose of the MT system.

After these questions, educators who responded to Q14 with "Yes" were directed to Block 3, which included questions Q18 to Q28.

4.1.2.3 Block 3 - MTQA Pedagogical Questions

In Q18, the responses of the participants shown in Figure 4.12.

The data indicate that HE is the most frequently taught method among TS educators, surpassing both automatic and semi-automatic approaches. In contrast,

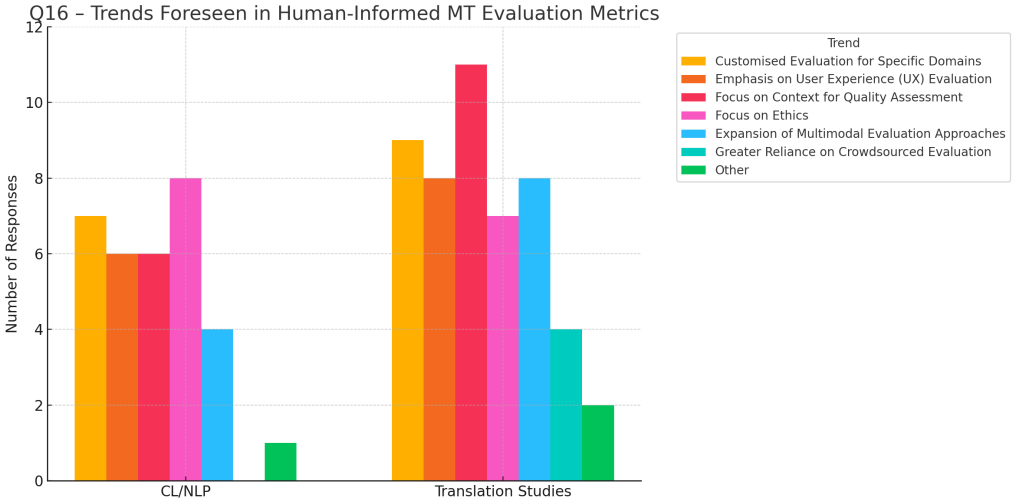


Figure 4.11: Trends foreseen in HE for MT, grouped by disciplinary background (CL/NLP and TS). Categories include customised evaluation for specific domains, user experience, context, ethics, multimodal evaluation, crowdsourced evaluation, and other. Multiple selections were allowed.

CL/NLP educators report teaching HE and AEMs equally often, suggesting a more balanced emphasis between the two. Semi-automatic methods are less frequently mentioned in both groups, with only minimal uptake. When asked for further comments in Q19, each group explained their emphasis. The CL/NLP participants explained how to teach MTQA as a complement to the main content, embedding quality assessment within broader courses. They report a balanced use of HE and AEMs, with references to BLEU, BERT, and COMET. They also mentioned that there is less emphasis on formal error annotation with frameworks such as MQM. TS educators, on the other hand, place a stronger emphasis on HE, especially with the use of DQF-MQM. There is pedagogical emphasis on post-editing, critical reflection, and the use of MT in real-world scenarios. AEMs are mentioned, but often discussed critically, such as BLEU being unreliable for certain language pairs, such as English-Turkish.

In Q20, when asked about whether they teach other NLP tasks, 18 participants responded with "No" and 1 participant responded "Yes". In Q21, when asked for further details, two responses mentioned that they focus entirely on translation technologies, and one response mentioned that they try to remind students that they can evaluate system performance using precision, recall and f-measure. Q22

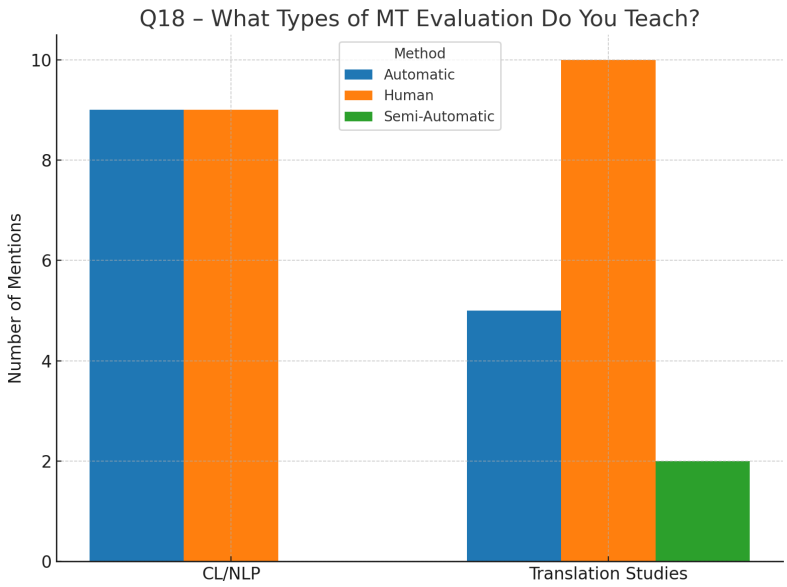


Figure 4.12: Types of MT evaluation methods taught, grouped by disciplinary background (CL/NLP and TS). Categories include Automatic, Human, and Semi-Automatic evaluation. Multiple selections were allowed.

asked about the length of time educators had been teaching MTQA, Table 4.2 shows the years of teaching per group.

Field	Minimum	Maximum	Mean	Std Deviation	Variance
CL/NLP	2.00	30.00	9.67	9.17	84.00
TS	2.00	20.00	7.70	5.85	34.23

Table 4.2: Descriptive statistics for *years* of experience teaching MT quality assessment, grouped by field

The average experience is slightly higher in the CL/NLP group (9.67 years) than in TS (7.70 years), though both groups show a comparable minimum and a reasonably concentrated distribution. The data suggest broadly similar levels of experience across the two fields.

Teachers were asked to assess the importance of teaching students how to plan evaluations for MT systems in Q23, and the responses can be seen in figure 4.13.

The data show a strong consensus between CL / NLP and TS on the importance of teaching MT evaluation design. All respondents rated it at least "Moderately Important", with the majority selecting "Very Important" or "Extremely Important". In CL/NLP, five educators rated it as very important and four as extremely important. In TS, five rated it extremely important, three very important, and two

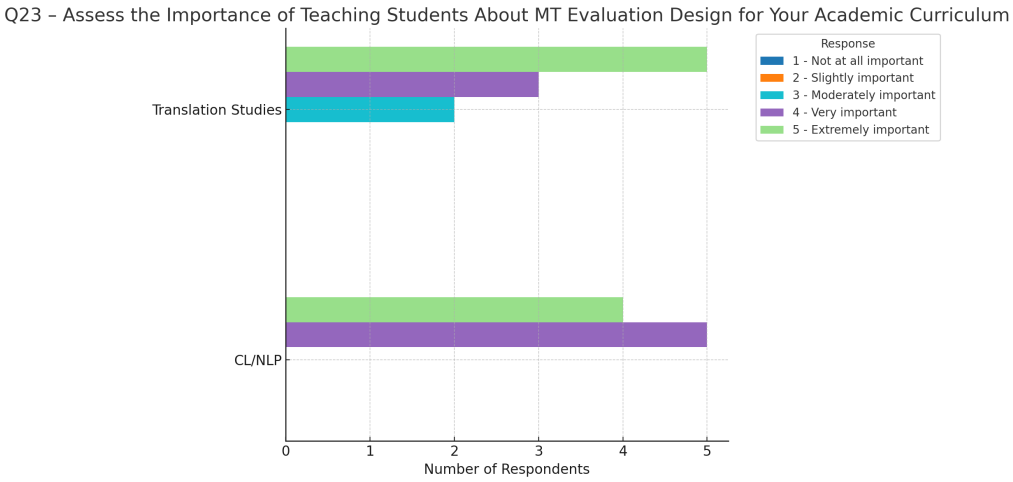


Figure 4.13: Perceived importance of teaching students about MT evaluation design, grouped by disciplinary background (CL/NLP and TS). Responses are shown on a 5-point Likert scale ranging from "Not at all important" to "Extremely important".

moderately important. When asked for further details in Q24, the responses included:

- Evaluation is a transferable skill, therefore it is useful.
- Evaluation leads to a better understanding of the model of the system, leading to more mastery over it.
- The planning of evaluation influence or compromise results.

In Q25, educators were asked to assess the importance of evaluation metrics, as seen in Figure 4.14.

The data show differing emphases on evaluation metric types across fields. In the CL/NLP group, responses for AEMs and HE are relatively balanced, with slightly higher ratings for HE at the top end of the scale. In contrast, TS respondents place greater importance on HE, with more selecting "Extremely Important", while AEMs received more varied responses, including lower importance ratings. When asked about further details on their choices (Q26), each group responded differently. The CL/NLP group explained:

- Technical courses should have a smaller focus on HE.

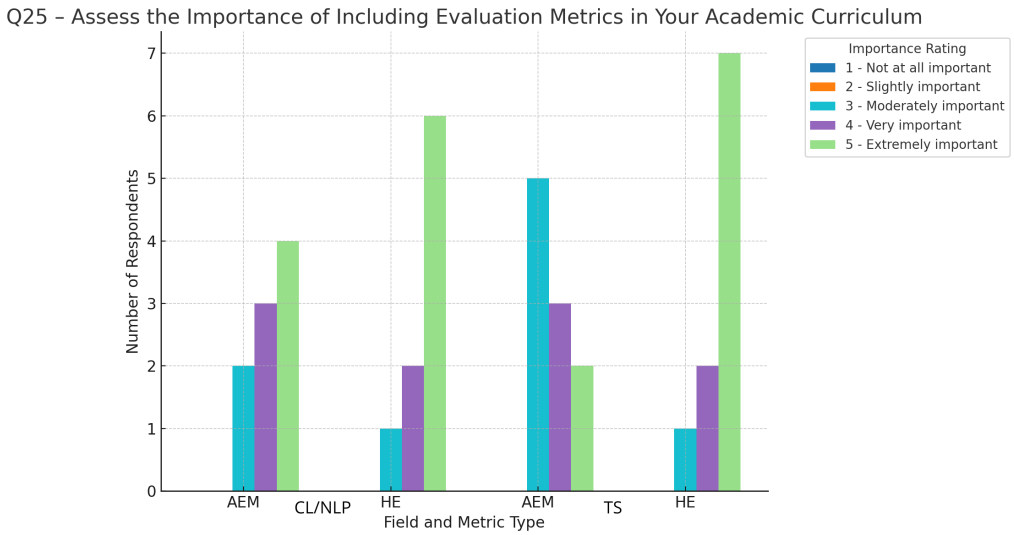


Figure 4.14: Importance of including evaluation metrics in the academic curriculum, grouped by AEMs and HE and disciplinary background (CL/NLP and TS). Responses are shown on a 5-point Likert scale ranging from not at all important to extremely important

- Depending on the type of student, there should be more focus on a certain type of metric.
- Because there is not much time in the curriculum, evaluation metrics may not be mentioned.
- AEMs are indispensable for development.

The TS group responded that both types are important, but depending on the type of student, you might want to focus more or less on a certain metric.

Noticing that content may not be the only difficulty in teaching MTQA, Q27 asked for pedagogical aspects that could be improved, as seen in Figure 4.15.

The data indicate that both CL/NLP and TS educators perceive similar pedagogical gaps in teaching MTQA, with emphasis on allocated hours. Curricular structure and faculty expertise were also recurring pedagogical issues. Notably, TS respondents reported equal concerns about allocated hours and the expertise of faculty when teaching MTQA. Scalability of teaching methods was mentioned only once per group, and no respondents selected "Other". When asked for further details in Q28, an educator from the NLP group reported difficulty organising group

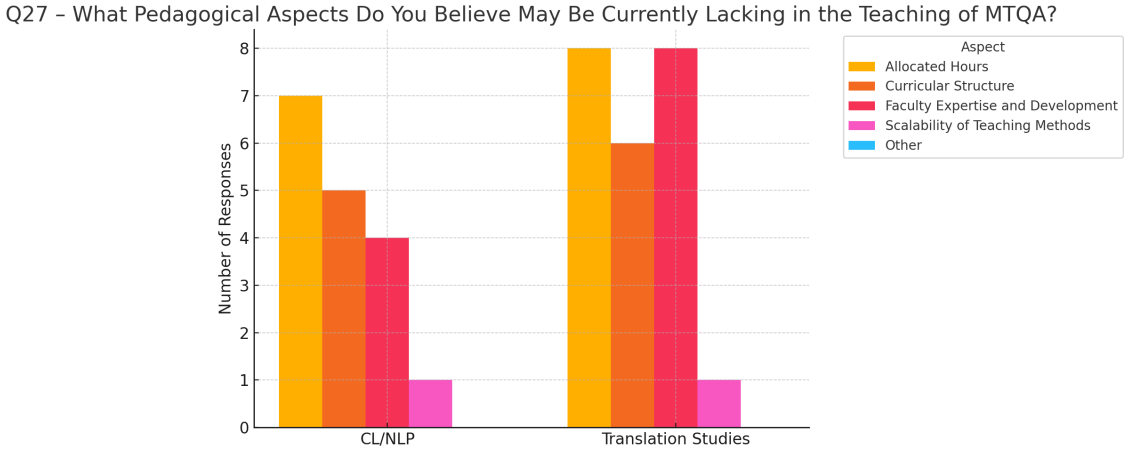


Figure 4.15: Pedagogical aspects perceived as lacking in the teaching of MT quality assessment, grouped by field (CL/NLP and TS). Respondents selected from: "Allocated Hours", "Curricular Structure", "Faculty Expertise" and "Development", "Scalability of Teaching Methods", and "other". Multiple selections were permitted.

work in laboratory sessions, as they had 40 students in their classes. TS teachers mention that there are not many pedagogical datasets available and they often lack the time to provide individual feedback to students.

4.1.2.4 Block 4 - Human Evaluation Teaching Recommendations

Q29 to Q33 were open to all educators again, even those who responded "No" to Q13. Q29 investigated what content could be useful for NLP students in a course, as seen in Figure 4.16.

From the NLP group, respondents prioritised "Adequacy/Fluency Measures", "Usability", and "Acceptability", with additional interest in "Comprehensibility", "Design of MT Evaluation", and "Error Analysis/Error Typology". "Translators as Expert Evaluators" and "Inter-Annotator Agreement" also received attention, while "Crowdsourcing for Evaluators" and "Ranking" were least selected. This suggests a concentration on metrics focused on system outputs, with moderate emphasis on methodological depth and professional roles. In contrast, the TS group reported broader and more evenly distributed preferences in nearly all categories. "Design of MT Evaluation" and "Translators as Expert Evaluators" received the highest selections, followed closely by "Adequacy/Fluency Measures", "Error Anal-

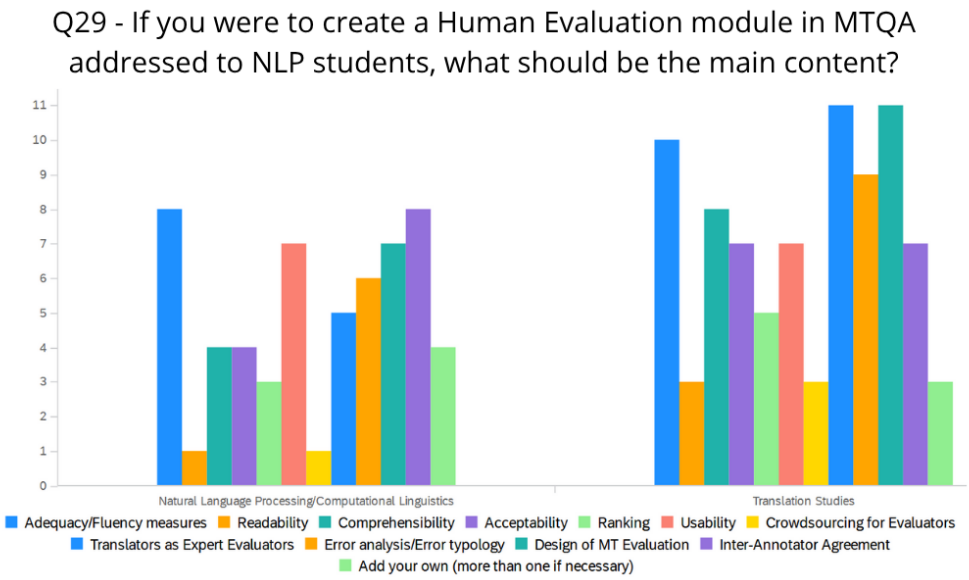


Figure 4.16: Suggested content for a Human Evaluation module in MT quality assessment, grouped by field (CL/NLP and Translation Studies). Responses reflect a range of potential content areas, with multiple selections being allowed.

ysis/Error Typology", and "Comprehensibility". "Acceptability", "Usability", and "Inter-Annotator Agreement" were also well represented. These patterns reflect a disciplinary inclination toward highlighting the strengths of translators as indispensable professionals and the methodological contributions made in the field.

When Q30 asked for more details, an educator suggested hands-on sessions incorporating these metrics. One educator suggested that objective measures are better than subjective measures. One educator suggested that error typologies are essential. Finally, one educator suggested that the components "Translators as Expert Evaluators", "IAA" and "Crowdsourcing" should be contained within "Design of MTQA", and not be taught separately.

Another question asking for suggestion, Q32, focused on the format of the training, as Figure 4.17 shows.

The data reveal clear contrasts in how each group envisions training delivery for MTQA. TS respondents overwhelmingly preferred formal academic integration, with limited interest in alternative formats such as short courses or workshops. In contrast, CL/NLP respondents demonstrated a broader range of preferences, indicating openness to more flexible and varied training formats. Workshops, multi-day short

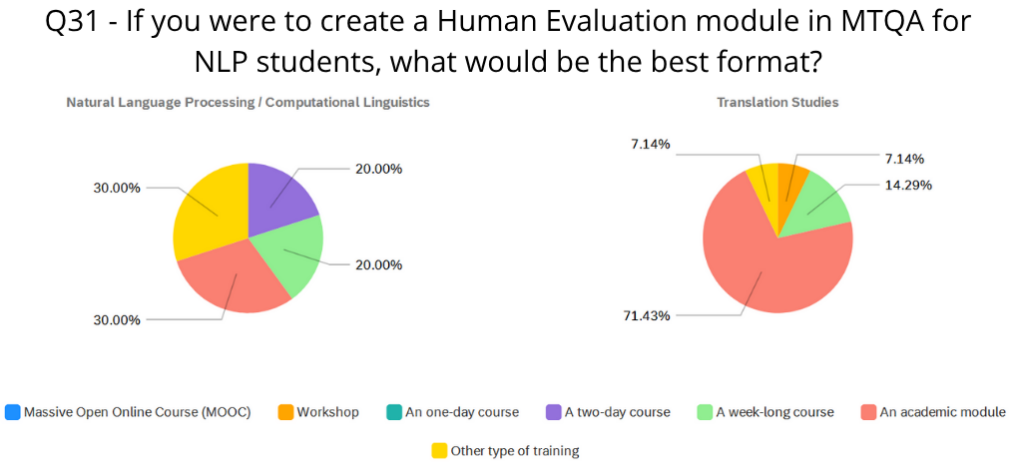


Figure 4.17: Preferred formats for training in MT quality assessment, grouped by field (CL/NLP and Translation Studies). Respondents selected from options such as workshops, short courses, academic modules, and online formats. Percentages reflect the proportion of selections within each group.

courses, and academic modules were selected with relatively equal interest. This pattern reflects a more modular or decentralised approach to incorporating MTQA into existing training structures. When asked to explain (Q33), one educator from NLP explained that a course lasting several weeks is needed, as students need to think and reflect with individual work, and another NLP educator suggested as many laboratory and hands-on sessions as possible. The TS group had several contributions, with one educator mentioning that the topic warrants an academic module; one educator's response reported that in-person is essential for troubleshooting with technical aspects; and one educator suggested spread-out hands-on sessions when it is not possible to have an academic module.

4.2 Phase 2 - Course Blueprints

A course blueprint is a preliminary inventory created prior to the development of a course (Dousay and Logan, 2012). The blueprint helps the ID professional link each component to create a syllabus and the course with its lessons, which can be seen in Figure 4.18.

In other words, this blueprint results in a syllabus and a course that follows

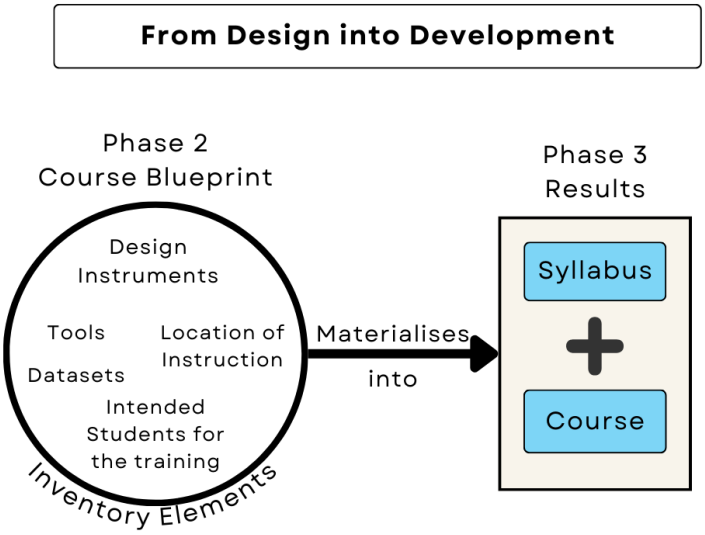


Figure 4.18: An illustration of how the course blueprint elements from Phase 2 - Design result in a syllabus and a course for Phase 3 - Development

strategies and tasks to meet the objectives of the course (Jonassen et al., 1999). This process is conceptualised by the term "coherence" regarding training, that is, training is designed and developed with guided materials and teaching methodology to result in more effective student learning (Wang et al., 2023). In the context of this thesis, I include recommendations from the literature where relevant. Thus, I believe that the following elements are important and must be closely related:

(a) Design Instruments, which serve as the orientation to support the decision making (Seel et al., 2017).

(b) Tools, which designate features of instruments, authoral or not, that support teaching and learning (Seel et al., 2017).

(c) Datasets, outlining how some material can be crafted to the learners' needs, in this context, curated datasets can be used to teach the subject matter, MTQA.

(d) Intended Participants, as the students' backgrounds and needs must be considered (Freire, 1996).

(e) Expected Physical Location of Instruction, a necessary component to define

due to possible constraints affecting development and implementation (Dick et al., 2021).

The **Design Instruments** organise the type of framework to plan the creation of the course. This structure consists of an instructional matrix (IM) and lesson plans. The **Tools** describe the type of technologies that will be adopted to create teaching materials. The **Datasets** explain the acquisition of data, its organisation, and how it is implemented into tasks. The **Intended Participants** refer to the type of student for the training, the recruitment process, and retention strategies within the context of the course. The **Expected Physical Location of Instruction** is a description of the intended site for the delivery of the course.

4.2.1 Design Instruments

The planning of the course started with the use of an IM. An IM can be defined as an overview of an entire course (Filatro and Bileski, 2017) that can be broken into smaller parts, such as units, to be reused in different contexts, distributed and shared with other teachers (Filatro, 2023). The components of an IM can be seen in Figure 4.19.

After the definition of the IM, sequenced lessons were designed for each unit. The sequenced lessons are inspired by the concept of "learning progression", which entails a sequence of lessons that are empirically validated (Duschl et al., 2011). In the context of this thesis, the concept provided by (Duschl et al., 2011) is useful for the dissemination of structured learning, not to understand learning itself. In addition, a requirement for understanding the proper progression of lessons is based on a needs analysis of the field (Duschl et al., 2011, p. 125), which was carried out in Phases 1a, 1b, and 1c (see Figure 1.1). Figure 4.20 shows the core ideas of the training, which helped to make the decisions about the content and how the lessons would progress. With these decisions, the organisation of the lessons was designed, as shown in Figure 4.21.

Each lesson plan contained a heading with basic information for documentation,

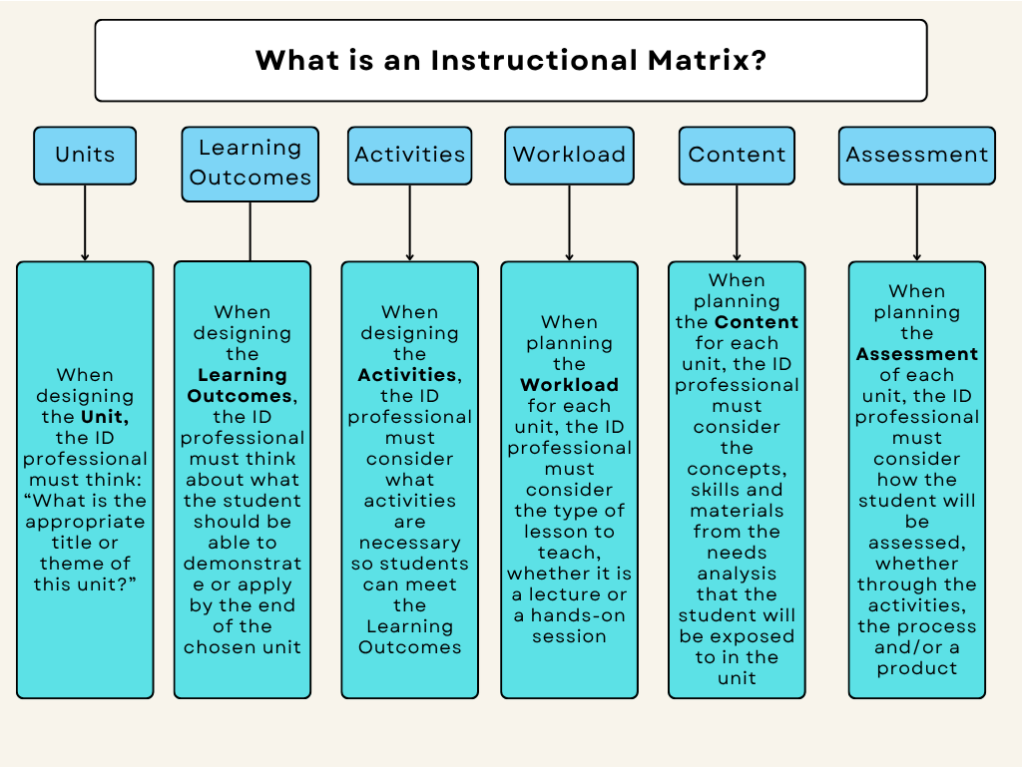


Figure 4.19: Definitions and directions of each element of an IM that the ID professional must follow in the design phase.

such as module name, lecturer, total learning time, and a description of the LO and the lesson that would be taught, as per Figure 4.22. Figure 4.23 shows an example of how the instructions are described for the teacher. In a table described as "Class Instructions" the teacher has access to three columns: The activity column describes the topic/task, name of the topic, and time taken by the topic. The methodological procedures describe how the teacher should proceed, whether the procedure is more teacher-centred or student-centred. And finally, the lesson plan also has the "tool" column, which describes the necessary tools for that specific activity, such as slides (and the respective number of the slide) or a specific software or resource to be used, such as a Jupyter Notebook.

The sequenced lessons accommodated the structure of moments where lessons would be teacher-centred with lectures and moments where lessons would be student-centred with lab sessions. These teaching styles are described in depth in Section 4.3.

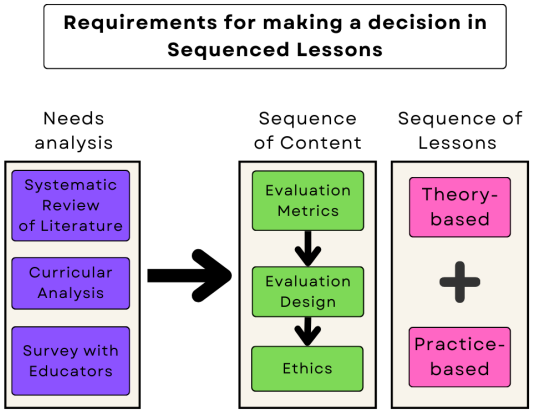


Figure 4.20: Requirements for showing how the needs analysis results in the sequencing of content and instruction structure

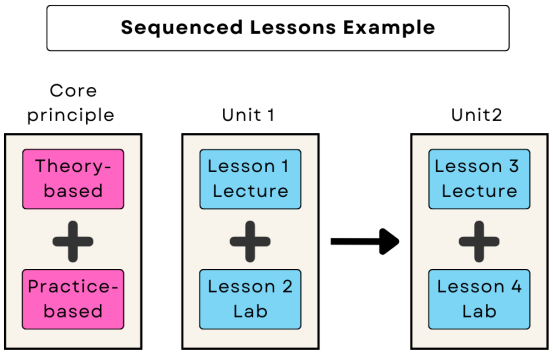


Figure 4.21: Rationale explaining the core principle of lesson structures and sequencing

4.2.2 Tools

The tools described in the next sections were all hosted in Google Classroom. The choice of this Learning Management System (LMS) was made for its easy setup, quick settings, and my familiarity with the tool. The LMS contained all materials and links that would direct students to the tools.

4.2.2.1 Tools for Lectures

This design blueprint foresaw that the course consisted of sequenced lessons for each unit (see Figure 4.23), a multitude of tools were adopted for several reasons. For lectures, tools that enhanced the visualisation of information and presented realistic

Module: Machine Translation Quality Assessment
Lecturer: João Lucas Cavalheiro Camargo
Lesson Plan: Lesson 1 out of x
Total Learning Time: Up to 2 hours
Description:

Learning Outcomes (LO)
LO1 - Develop and Foster the awareness of the role of HE in the development of MT systems.
LO2 - Develop and Foster the awareness of the role of AEMs in the development of MT systems.
LO3 - Develop and Foster the understanding of key elements in the development of MT systems, including purpose of system, evaluation metrics, type of evaluators and ethical considerations, to ensure systems are less risky and less biased.

Figure 4.22: Heading of the Lesson Plan detailing basic descriptors of the lesson before the instructions

CLASS INSTRUCTIONS		
ACTIVITY	**METHODOLOGICAL PROCEDURES**	**TOOLS**
Topic 1 Warming Up 5 minutes	The teacher will begin the lesson asking students how they are doing. The teacher will ask students about the reflective assignment from the previous lab, in case any student had not finished at the lab.	-
Topic 2 Has Machine Translation	The teacher will proceed with the lesson, and the teacher will open the slides. As a warm up, the teacher will ask students what they know about the concept of "human parity". Mention to students about studies that claimed	Slides 4 to 12

Figure 4.23: Heading of a Lesson Plan detailing basic descriptors of the lesson before the instructions

situations were chosen[4]. The characters would participate in conversations to answer questions or introduce a new topic to be discussed in the lecture, as can be seen in Figure 4.24.

Having relatable characters to create a story to teach certain topics can be beneficial in STEM education, as suggested by Cao et al. (2023). This effort was made because there is a need to further increase the diversity of representation in educational materials (Seburn, 2021), especially because STEM education requires more interventions in this regard, adding to ongoing academic efforts to change the status quo, that is white and male dominated (Meyer et al., 2025). For example, the slide in Figure 4.25 was designed to explain the concept of IAA. It contains small boxes

[4]For example, a combination of Canva and ChatGPT (versions 3, 4 and 4o) were used. Canva was used for elements that would enhance the visualisation with figures, scenarios and customisable slides, while ChatGPT was used to generate cartoon-like characters.

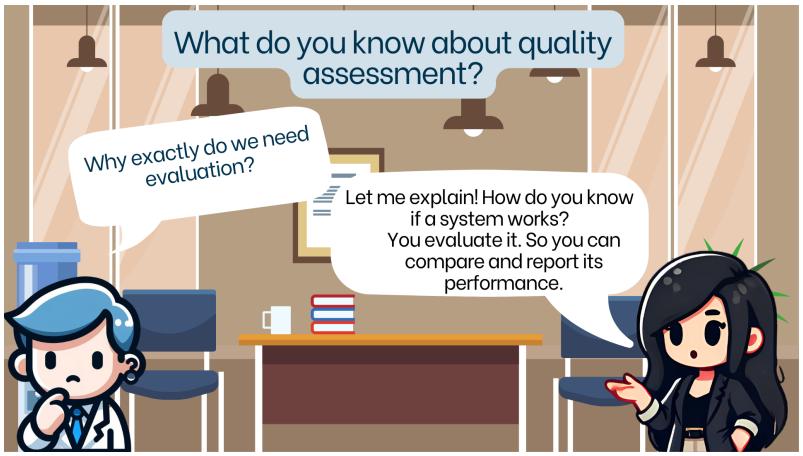


Figure 4.24: Slide illustrating a conversation between a blue-haired character and a mentor-like character talking about the need for measuring the quality of MT systems

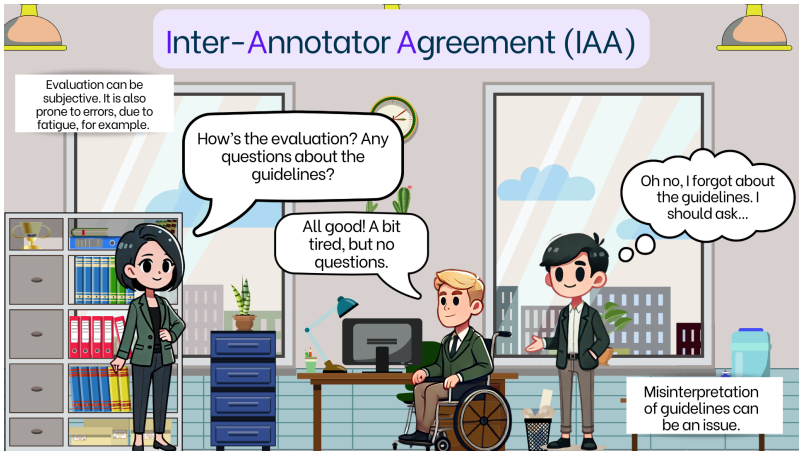


Figure 4.25: Slide showing a scenario of evaluation with different characters, including a woman character as a team lead, a scientist with a disability and a non-binary scientist.

with useful information about measuring IAA, showing characters in situations that warn students about potential challenges.

Vevox was chosen for interactive quizzes[5], with questions formulated using different formas. Figure 4.26 shows multiple choice, scales, and questions that show responses as word clouds.

It can be used to check students' understanding of topics covered in previous lectures, to allow students to perform self-assessment and respond to questions to form word clouds.

[5]Dublin City University (DCU) has access to the interactive polling tool Vevox, made available by the Teaching Enhancement Unit. This polling tool allows the teacher to design questions that can be projected on a screen. Students can answer these questions using their phones.

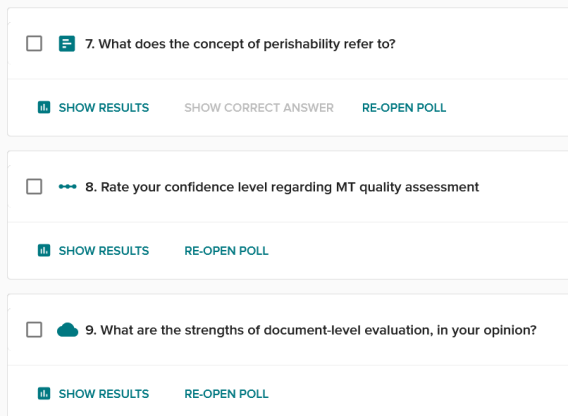


Figure 4.26: Interface of Vevox, where the teacher can create multiple choice questions, scale-based questions and word clouds.

4.2.2.2 Tools for Labs

The tools for lab sessions were more interactive, as they were supposed to be student-centred. Hence, the following tools were selected:

(a) Google Sheets

(b) Qualtrics

(c) Jupyter Notebooks

Google Sheets can be useful for organising tabs with different tasks, showing sentences in the original language and the MT output. It allows the use of drop-down menus to select different options. **Qualtrics** offers similar options, but its UI layout adapts to the use of computers and phones, and a more customisable back-end allows for a more friendly user interface. This contrast in tools, such as Google Sheets and Qualtrics, can be used for the same type of tasks, but offering different experiences, which can be a point of discussion in matters of MTQA. As seen from the SRL (Subsection 3.5.1.2), **Jupyter Notebooks** are commonly used in TS for the teaching of MTQA, and because of its positive results, are included in the list of tools used.

4.2.3 Datasets

In Phase 1c (Section 4.1), one of the teachers' suggestions was the use of pedagogical datasets. Therefore, a corpus in which "pedagogical value" is added to documents for certain tasks (Sheng et al., 2017) was compiled with a set of issues for the students to analyse (Bart et al., 2018). This dataset demonstrates specific cases where a certain type of issue would be required to be analysed in MTQA processes.

4.2.3.1 Dataset and Annotation Criteria

The methodology of corpus compilation followed the strategy used by Castilho et al. (2021) for the DELA Corpus, where the authors looked at challenging texts where MT systems would have difficulty translating due to certain issues, such as ambiguity of lexical or gender nature, identified in Castilho et al. (2020). Some examples were adopted from the DELA Corpus with inverted language direction (Brazilian Portuguese into English). Examples were adapted to ensure the MT system outputs would generate disfluent output in English. The goal was to show the errors of NMT systems from the perspective of monolingual users, users who do not know the source language and those who only see the MT output.

4.2.3.2 Using the Dataset for Pedagogical Purposes

I compiled the dataset for different evaluation scenarios with distinct evaluators: i) translators: texts may be used to measure adequacy and fluency; ii) end users: texts may be used to understand a product or location review, focusing on comprehensibility. For the translator's perspective, I compiled a dataset of six different texts. [6] The domains for the six texts were:

- Text I - Literature
- Text II - User Generated Content (Social Media Post)
- Text III - User Generated Content (Product Review)

[6]The dataset is available on `https://github.com/joao-cavalheirocamargo/MTQATraining/`

- Text IV - User Generated Content (Product Review)
- Text V - User Generated Content (Recipe)
- Text VI - Academic Abstract

These texts were chosen from a variety of sources, primarily because they were freely available online. They were compiled for the sole purpose of measuring their fluency (focusing on the Tt), because participants were not expected to speak Brazilian Portuguese (the original language). The texts compiled for use by end users consisted of product reviews and location reviews, acquired from Amazon and Google Maps. These texts were chosen so that students may answer the question "With this MT output, can I understand this product/location/restaurant review?". That is because it might have been a question they have asked themselves as end users. These texts were later used to design a task focused on comprehensibility (see Tasks 1 and 3 of Subsection 5.2.2 in the next chapter).

4.2.4 Participants

In educational research, a teacher can opt to conduct research with their students or students to whom they have access for small-scale testing (Cohen et al., 2018). I selected **master's students in NLP**, because I had access to them. To formalise this, the following elements were considered: i) type of student; ii) access and; iii) inclusion ethics. When selecting criteria for the **type of student**, master's NLP students were considered essential to this research. Moreover, participants with similar curricular background were also included in the recruitment for the course, to avoid having a low number of participants. That is the case of a specific type of purposive sampling, which is called homogeneous sampling (Cohen et al., 2018). Homogeneous sampling allows the researcher to consider students with a similar profile to master's NLP students, therefore PhD students in the field of NLP and M.Sc students of Translation Technology were also included. Although not as

technical as NLP students, translation students bring a different perspective and can benefit from studying MTQA from the perspective of a developer.

Considering how university students can prioritise their own studies and the duration of study can impact retention in research scenarios (Khatamian Far, 2018), certain **access strategies** had to be applied in recruitment, which consisted of two phases. In the first phase, I asked the lecturers of the respective programmes in DCU for permission to advertise the course in one of their lectures. I described all aspects of the course, from the objectives to the dates, locations, and recommendations on how to balance my optional training course with their mandatory modules. Considering the duration of the course and the potential workload were important factors, I assured students that the training course would not require any homework, and that any assignment or activities could fully be completed in the hours of the course. Then I provided instructions on how to sign up for the course, including my email for further questions. For the second phase, a recruitment email was sent to members of the research centre I belong to via email. During the recruitment, students were reassured that the training would not be related to their ongoing modules and that they should not feel pressured to participate in the training. Moreover, they could withdraw from training and study at any time (following **inclusion ethics**, an aspect relevant in educational studies (Liebel and Chakraborty, 2021)).

Twenty students participated. Although the number is not enough to achieve statistical power (Cohen et al., 2018), the low number of participants supports an in-depth qualitative investigation (Creswell, 2015). The qualitative focus, then, allows for a supplementation through classroom observation (discussed in Subsection 4.4.3).

4.2.5 Physical Location

Physical locations directly affect the pedagogical approaches implemented in the classroom (Monahan, 2002), as certain classroom layouts can enhance student engagement (Rands and Gansemer-Topf, 2017). Due to this, a lab facility of the School of Applied Language and Intercultural Studies at DCU was chosen.

The room has four pods with five seats and five computers each. Hence, students are not in rows but are seated in groups. The room has projectors and two screens opposite each other, on each wall. This is particularly useful when visual aids are required, the teacher can connect their laptop with VGA and HDMI ports. These elements provided the necessary resources to conduct the theory-based and practice-based classes (see Section 4.2.2.1). For practice-based classes, students are already in groups, although there is nothing to stop the teacher from reassigning groups when necessary. The computers can be used freely by students, provided that they are DCU students. As the participants described in Section 4.2.4 are primarily DCU students, they can freely use the computers. However, foreseeing that participants from the ADAPT Centre were also invited, students were advised to bring a laptop to be able to engage with the materials.

4.3 Phase 4 - Syllabus Implementation Methodology

This phase succeeds Phase 3, discussed in detail in Section 5.1. The methodology of Phase 4 includes the description of the teaching methods for the delivery of the type of course described in Phase 2 (Section 4.2). While the curricular analysis (Section 3.7.3) was unable to reveal any specific pedagogical procedures, focusing more on content, the explicit discussion of teaching methodology in NLP classes is important. Fei et al. (2024) states that in STEM education there is still more potential to address pedagogical barriers.

To this end, this section describes the different methodologies that were considered for the development of the courses, specifically, the chosen teaching methodologies that impact how the course is developed, from instruction to tasks. In addition, this section is influenced by the discussion about educational philosophy in Section 2.2.2, since the approach that I adopted to develop the teaching methodology is influenced by critical pedagogy (Freire, 1996).

4.3.1 Theory-based classes - Lectures

Lectures are methodological procedures in which the teacher uses a didactic language to present, define, analyse, and explain themes and topics in a course (Mattos, 1976, p. 187). Although lectures are one of the most traditional approaches to teaching in higher education (Tronchoni et al., 2022), studies generally do not elaborate on teaching methodology (Tronchoni et al., 2022, p. 13). This section aims to define, describe, and reflect my view of teaching methodology to be used in this thesis. Andreata (2019) distinguishes between the two directions of a lecture, the traditional type and the dialogic type, which is exemplified by Figure 4.27.

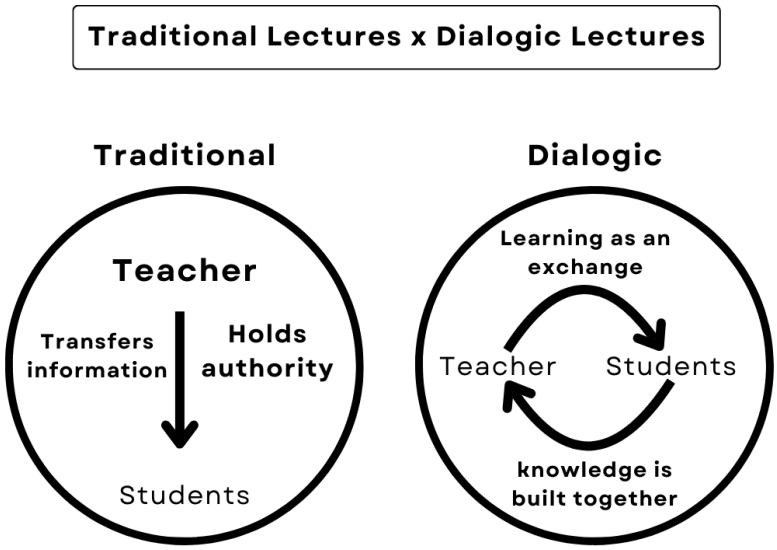


Figure 4.27: Differences between traditional and dialogic lectures, with the traditional showing the teacher in a vertical relationship with students, while the dialogic shows a horizontal relationship, where knowledge is built together

Traditional lecturing assumes that the teacher holds all the knowledge and simply transfers information to students. The limitations of this methodology are that it can position students in a passive position, depending on their interest and attention, and it can also lead to fatigue (Mattos, 1976). In the other teaching orientation, dialogic lecturing, the concept "Dialogic" can be applied in different contexts, such as: an ontological orientation, which informs students about what is being studied; an epistemological orientation, which guides how a teacher constructs knowledge with their students; or the act of dialoging (Mercer et al., 2023). I adopt the con-

cept of "Dialogic" both ontologically and epistemologically. Informed by ontology, I view humans as historical and social figures, making the humane aspect central in learning, as humans reflect, interpret and act upon the world. From an epistemological standpoint, I position myself, my students and our shared dialogue as the central means of producing knowledge for this research. This is how I approach **dialogic lecturing** by embedding Paulo Freire's principles into practice (Andreata, 2019). It is a method of teaching that has evidence to impact the success of student outcomes (Hardman, 2020). Thus, the elements that define a dialogic lecture are adopted from the research of Andreata (2019, p. 719):

(a) Use dialogues with students.

(b) Adopt ethics as a factor in teaching.

(c) Respect students' previous knowledge.

(d) Foster students' autonomy.

(e) Foster students' critical thinking through questions.

For **(a) using dialogues with students** the use of materials that foster dialogues or the visualisation of dialogues (Figures 4.24 and 4.25) helps students understand that it is perfectly acceptable to ask any type of question. It particularly helps if the teacher themselves appears on the material. **(b) Adopting ethics as a factor in teaching** is addressed by adding "Ethics" as a content of the training (See Section 5.5 of Chapter 5). **(c) Respecting students' previous knowledge** is directly transferable if a student uses previous knowledge of another NLP task to MTQA or vice-versa. **(d) To explicitly foster students' autonomy**, the teacher must suggest that any type of question and interaction are welcomed during the classes and in terms of course content, that students should experiment with the resources from the classes (studies, tools, metrics, etc.) not just within the class, but in their own practice. Lastly, to promote **(e) critical thinking of students through questions**, I follow the example of McCrae (2024) to adopt a co-creational

stance, where at every instance of the course, students may suggest, ask about or critique topics covered (or not) in the lecture. For example, the student may critique the usefulness of a tool or a practical aspect of the course, particularly if they refer to the literature to support their critique or comment.

4.3.1.1 Role of the Teacher and the Students

My role in dialogic lessons is not to serve as a teacher who holds all the information and who deposits information in students' minds, but to act as an educator who facilitates structured knowledge and activities to learn and reflect with students, following Freire (1996). The discussions are guided by slides where they show the dialogues I had with my supervisors to show students my thoughts, attitudes, and questions when I first started studying MTQA. The structure of the lessons in this way allows the content to be shown in a more authentic setting. It helps students see themselves as learners without the stigma of "not knowing enough". This approach shows students that, as a teacher, I am not someone who always knew or knows everything.

In this way, the role of the students is to experience, interact, and be more or less active at different moments of the lesson. As a feature of dialogic teaching, students were encouraged to bring their previous knowledge from other areas of NLP.[7] Furthermore, it is expected of students to speak confidently throughout the course of the training, as this can become part of classroom culture(Cui and Teo, 2020, p. 12).

4.3.1.2 Duration of the Classes

The lessons were planned to last one hour, but extra time was allotted for any particular topic that may have garnered more interest, for possible discussions/questions. Pedagogically, students were also allowed to finish exercises that were not completed in a previous lesson. Imagine the scenario where in the second theoretical lesson (les-

[7]The course also had students from translation enrolled, so their value in interacting and dialoguing provided more nuance and depth in linguistic analyses.

son 3), if students had not finished tasks from the first practice-based lesson (lesson 2), they could use the remaining time to complete exercises and ask questions. Because this is an optional course, this strategy allows me to use any extra time as a retention strategy. Therefore, two hours per lecture were scheduled.

4.3.2 Practice-based classes - Lab sessions

After being instructed on theory in lessons, students have labs to put into practice what they learned. The labs sessions were student-centred, closely tied to what is known as active learning, in which students can construct knowledge through activities and tasks (Doolittle et al., 2023). I also aimed for collaborative environments, where students could talk and discuss, unless there was a specific reason for them not to do so. This follows from the idea that successful practical classes in Computer Science can occur in small groups in collaborative settings, as Córdova-Esparza et al. (2024) suggests. The use of active components results in more acceptance and positive feedback from students (Berssanette and de Francisco, 2021).

The choice of using classroom activities comes from studies in TS, outlined in the SRL (Section 3.5.1.4 of Chapter 3), which succeed when employing task-based learning (Esqueda, 2021). In computer science, a combination of tasks is usually adopted, aligning the approach more with problem-based learning (Barg et al., 2000), which is an experiential learning approach around a specific problem that requires specific tasks to address it (Hmelo-Silver, 2004).

4.3.2.1 Role of the Teacher and the Students

For practice-based classes, the teacher takes a step back and becomes less of a central figure and more of a mentor, guiding students. The teacher poses realistic scenarios where the students have to engage, interact, and make meaning out of their experience through active learning. When teaching about evaluation metrics, students should be provided with tasks about the practice of several HE metrics through the perspectives of different stakeholders, such as end users or translators.

Or, in the case of ethics, students were presented with issues in MTQA, such as claims that MT can substitute humans. Students had to discuss how to correct those exaggerated claims. If students did not know how to plan an evaluation, they could be presented with a scenario in which they must adequately plan how to evaluate an MT system. These scenarios pose realistic problems that can be structured with tasks that students can actively engage in, building on their dialogic lessons from the theory-based classes. They are able to critically examine, discuss, and collaborate with colleagues with such a design choice.

4.3.2.2 Duration of the Classes

A two-hour lab session was assigned for the practice-based lessons. Several tasks were designed around a problem for each session. Realistically, the tasks should occupy one hour and thirty minutes and the remaining time should be reserved for discussions and/or reflective reports. The deliberate choice to have time for discussions is a feature of problem-based learning (Hmelo-Silver, 2004, p. 237) and is in line with the features of dialogic lectures. The reflective component is a recommended principle derived from the SRL (Section 3.6), which means that students need enough time to think about the tasks they solved and how they addressed a problem in MTQA.

4.4 Phase 5 - Evaluation

The final stage, Phase 5, aims to assess the effectiveness of the course. Figure 4.1 shows multiple steps defined for the Evaluation phase, according to the multiphase mixed methods design (Creswell and Clark, 2017). It combines the idea of a one-group time series (explained in Subsection 4.4.2) and the analysis of the course through classroom observation (Subsection 4.4.3. As there are different data sources in different paradigms, the appropriate choice to explore, explain or complement one source to another is to use **Triangulation** as an approach to analyse results.

4.4.1 Triangulation

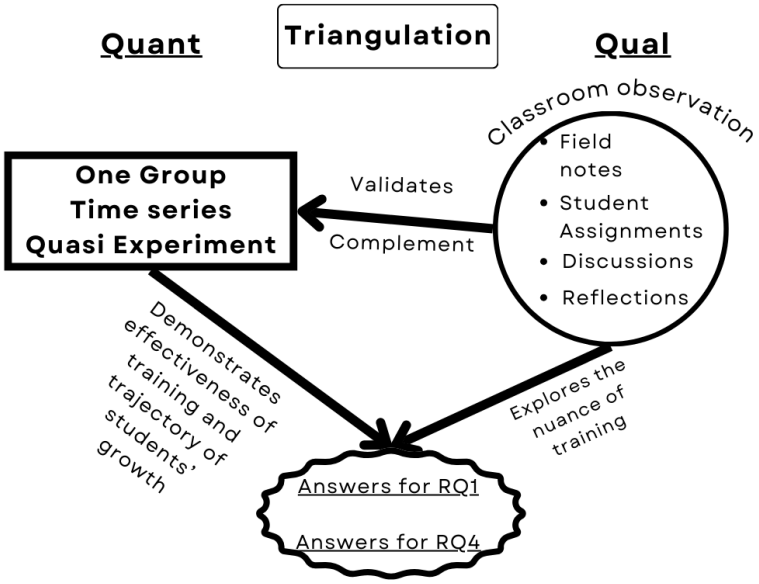


Figure 4.28: Figure illustrating the integration of a One-Group Time Series Quasi-Experiment and Classroom Observations to examine training effectiveness and students' confidence progression in MTQA, where qualitative data validates and complements quantitative findings to answer RQ1 and RQ4

Triangulation (Figure 4.28) is an approach in mixed methods research that adopts different perspectives in data collection and analysis to attempt to address the complex dimensions that participants may demonstrate in research (Cohen et al., 2018). As Mathison (1988) suggests, adopting a triangulation strategy allows the researcher to fully understand why the data may converge, be inconsistent, or contradictory. The strategy of adopting classroom observation aims to enhance the one-group time series experiment, as qualitative data can be used to "enrich" or interpret quantitative data (Rossman and Wilson, 1985). In this case, I will adopt the strategy of methodological triangulation, which has the assumption that combining multiple methods is a good possibility to respond to science inquiries, in this case, to respond to my RQs (Heesen et al., 2019).

The purpose of mixed methods research is to use specific analysis procedures to interpret and validate the data and results (Creswell, 2015, p. 254). The quantitative component of Phase 5 may be seen as an empirical validation, which, when related to the qualitative component, can be used to check the validity of the convergence of

these distinct sources, an approach recommended by (Creswell and Creswell, 2023, p. 51). That is, the results of the measurements should be accurate if they are consistent, and the qualitative component also adds depth to the findings.

4.4.2 One-Group Time Series

For the quantitative component, an one-group time series experiment was chosen. This is an experiment in which the researcher measures a variable at several time points in an experiment (Cohen et al., 2018). In the context of this research, it is a quasi-experiment. This means that the researcher does not randomly assign participants for an experiment, as is typical feature of empirical studies in educational research (Creswell, 2015), as randomisation or control is not always expected in education (Cohen et al., 2018). The stages are planned according to figure 4.29.

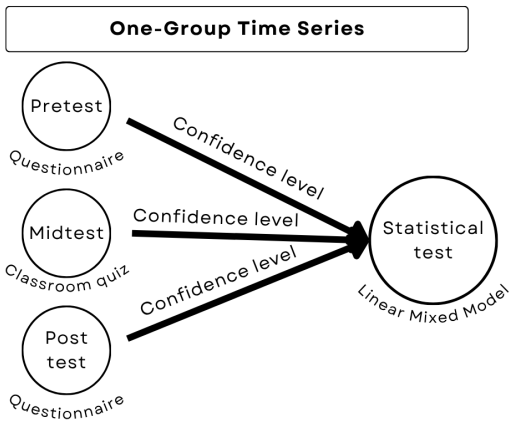


Figure 4.29: Figure illustrating the phases of the One-Group Time Series Quasi-Experiment showing the pretest, midtest and posttest converging the measurements into statistical results

This type of experiment is often used to report the trends of effects (or lack of) when implementing a new teaching method or curriculum for a group (Cohen et al., 2018, p. 407). It can provide insight into the progress of students over the course. The insights can vary, as they can be related to the core curriculum or personal matters (Ary et al., 2010). The pretest offers a baseline, the midtest integrates routine activity in data collection, and the posttest measures the final changes in characteristics observed throughout the training (Creswell, 2015). Despite the use

of this experimental design, certain limitations remain, primarily the absence of a control group, which restricts the ability to isolate the training's effect (see Section 4.4.3 for further elaboration on these challenges and how they can be overcome with a qualitative approach).

4.4.2.1 Pretest Questionnaire and Procedures

The pretest questionnaire (Appendix G) was administered during the first lesson of the training. The instructor introduced the research, invited the students to participate, and projected a QR code using the two projectors in the room (Subsection 4.2.5). This QR code linked to a Qualtrics survey, where students completed the pretest questionnaire, which contained the following types of questions:

- **Personal:** Questions related to individual information (Q1, Q2, Q3)
- **Attitudinal:** Questions related to opinions and perspectives on MT evaluation (Q4, Q5, Q6, Q7, Q8)
- **Behavioural:** Questions assessing confidence in using evaluation metrics and prior experience with evaluation (Q9, Q10)

Personal questions served to categorise students, pseudonymise their data, and facilitate grouping. Attitudinal questions explored students' perspectives on evaluation metrics, primarily through open-ended responses. Behavioural questions provided a baseline measure of students' confidence levels regarding different metric types. These questions used Likert scales (Likert, 1932), a widely used method in educational research to measure self-reported confidence and perceived competence (Ary et al., 2010). Measuring both attitudinal and behavioural aspects allowed for a more comprehensive assessment of prior knowledge and perceptions of students. Furthermore, open-ended attitudinal questions (Q4, Q7, and Q8) provided deeper insights. Q4, in particular, was valuable in identifying student expectations, helping to assess whether these expectations were aligned with the content and results of the course.

4.4.2.2 Mid-test - Classroom quiz

Intermediate assessments should ideally be integrated into regular classroom activities to avoid disrupting learning (Ary et al., 2010). The benefit of this mid-test is to measure whether the training could have already benefited students, as it happens in times-series designs (Gravetter and Forzano, 2014). Therefore, a classroom quiz was selected as the mid-test measure (see Figure 4.26), aligning with recommendations that intermediate assessments should integrate naturally into the course structure. The quiz was designed using Vevox (described in Section 4.2.2.1) and focused on the material of the first two lessons, assessing the students' understanding of MTQA metrics. It contained multiple-choice questions followed by additional exercises and discussions. This sequencing ensured that Likert scale-based behavioural questions did not stand out excessively, maintaining the coherence of the quiz within the course structure.

4.4.2.3 Posttest Questionnaire and Procedures

The posttest questionnaire (Appendix H) was administered during the final training session. Unlike the pretest, the posttest was distributed via a Google Classroom link, allowing students to complete it on their laptops or mobile devices. The questionnaire included the following question types:

- **Personal:** Questions related to individual information (Q1).
- **Attitudinal:** Questions related to opinions and perspectives on MT evaluation (Q2, Q3, Q4).
- **Behavioural:** Questions assessing confidence levels and experience in MTQA (Q5, Q6, Q7, Q8, Q9, Q10).

The attitudinal questions aimed to assess changes in students' perceptions of evaluation metrics, with a particular focus on ethical and technical aspects. The behavioural questions evaluated the students' confidence in applying the metrics,

alongside a newly introduced component: their ability to plan evaluations. Comparing pretest and posttest results allowed for an analysis of learning progression.

In addition, open-ended questions provided qualitative insights that complemented the quantitative results. These responses helped identify specific areas where the training failed or was not successful. In particular, Q8, Q9, and Q10 encouraged students to reflect on their learning experiences, providing additional context to interpret the quantitative findings.

4.4.2.4 Statistical Analysis

Two types of variables can be identified: independent and dependent variables. Independent variables are those that are manipulated by the researcher (Ary et al., 2010), which in this study refers to training. The dependent variables are those affected by the independent variables (Cohen et al., 2018). To represent this, Figure 4.30 was created.

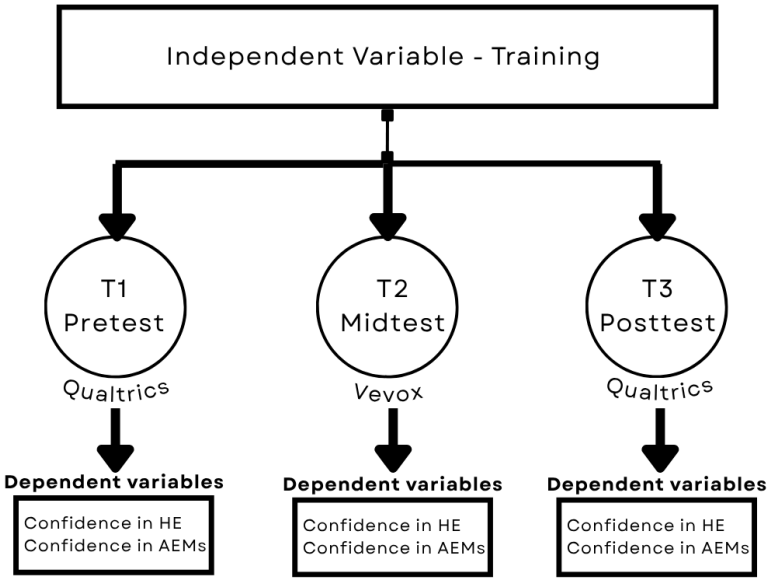


Figure 4.30: Variables of the study. The independent variable is the training intervention, operationalised as three timepoints (T1–T3). At each point, confidence in HE and AEMs was measured.

These variables were analysed using a Linear Mixed Model (LMM)[8] (Walker et al., 2019). This statistical test was chosen because it handles small samples and

[8]Calculated on Statistical Package for the Social Sciences (SPSS).

is able to track individual trajectories (Ker, 2014). It also has the ability to compute missing data (Muhammad, 2023). The background of the students (NLP or TT) was also calculated as a covariate. Its inclusion serves to adjust the fixed effects for individual differences in field and helps isolate the effect of the training, but it was not a primary independent variable. The LMM is thus represented in the following equation:

$$\text{Score}_{i,j} = \beta_0 + \beta_1 \text{Time}_j + \beta_2 \text{Background}_i + \beta_3 (\text{Time}_j \times \text{Background}_i) + u_i + \epsilon_{i,j} \quad (4.1)$$

Where:

- $\text{Score}_{i,j}$ represents the score of individual i at time point j.
- β_0 is the fixed intercept, representing the expected score at baseline (pretest) for the reference group (NLP students).
- β_1 is the fixed effect of Time, indicating the average change in score over time.
- β_2 is the fixed effect of Background (NLP vs TT).
- β_3 represents the interaction effect between time and background, determining whether students from different backgrounds show different learning trajectories.
- u_i is the random intercept, which captures individual-level variations in baseline scores.
- $\epsilon_{i,j}$ is the residual error term, representing unexplained variability at each time point.

Each β coefficient has a specific purpose in the equation, measuring effect size, being able to quantify the relationship between the predictors (time, background) and the dependent variables. β indicates effect sizes, but it does so through an

estimated change in the dependent variable for each measurement (Cohen et al., 2018, p. 805). β_0 serves as a baseline score to measure students, while β_1 is able to measure the average change in scores for each time a measure is taken. The focus of the study is to examine the progress of NLP students, but other groups were included as an additional predictor and a logistical possibility, leading to the use of β_2 as another element of the equation. It acknowledges the possibility of other backgrounds, treating this particular feature as an independent variable as well. Then, β_3 examines the relation between the background and time, to verify possible changes in how they learn. u_i is a feature of the LMM, particularly for how each individual within the group can change, as it is used in education to investigate students within certain hierarchies, such as cohorts, schools and other characteristics (Cohen et al., 2018, p.837). Finally, $\epsilon_{i,j}$ is included in the equation to account for fluctuations within individuals that arise due to random errors of measurement, which contribute to unexplained variability in scores (Ary et al., 2010, p. 254).

4.4.3 Classroom Observation

As outlined in Figure 4.28, classroom observation was used to complement the one-group time series experiment. Observation is when the teacher takes notes and reflects on their own practice, the students, the students' learning, and other aspects the teacher may choose (Freeman, 1998). In this research, I was the teacher and researcher, so I conducted the observation and took notes of my own lessons. To accomplish this, teachers can adopt a positioning named "teacher-researcher" (Bortoni-Ricardo, 2008), from participative educational ethnography, where the unification of the labels "teacher" and "researcher" occurs, to verify their synergy in action, being reflective, positioning teachers as creators of knowledge (Freeman, 1998, p. 2). When implementing a new type of training or content in the classroom, the teacher can be in a state of mind that reflects curiosity, leading to exploring alternative explanations for the phenomena that occur in educational contexts (Freeman, 1998, p. 36). While the "product" is valuable as an output of research in education,

a "teacher-researcher" finds great value in studying the process (Bortoni-Ricardo, 2008, p. 41). This frame views the roles of "teacher" and "researcher" in a dialogical manner, one contributing to the other (Cochran-Smith and Lytle, 2009). This framing is aligned with my pedagogical framework rooted in Paulo Freire's work (Freire, 1996). Most importantly, as a teacher-researcher it is possible to compare theory and practice, to reflect and understand praxis, to improve my own practice, and inspire other teaching practices in NLP. A discussion of the results of the implementation of the course from this perspective can be found in Chapter 6.

To achieve this, it is suggested that the teacher-researcher should adopt tools in their data collection that do not deviate much from classroom practice, through field notes or "research notes" (Bortoni-Ricardo, 2008, p. 46). These notes are important because they can register the complex reality of the classroom, and this reality is generally impacted by immediate actions, not leaving much room for complex thought (Freeman, 1998). However, most importantly, these data collection tools are not neutral. Data collected through this manner is filtered by the teacher-researcher's perceptions, thus, data is "constructed" by their interpretation (Lankshear and Knobel, 2004, p. 172). This is where the pedagogical rationale of combining TH (presented in section 2.2.1.2) and Freirean pedagogy (presented in section 2.2.2) is pertinent. Translation happens through active interpretation, because language happens through active interpretation. Thus, the subjectivity of observations by a teacher-researcher is not random, but grounded in the reality as per elements found in the classroom. Such elements can be classified as **observed data**, **written data** and **spoken data** (Lankshear and Knobel, 2004).

As noted above by Bortoni-Ricardo (2008), observed data is the systematic observation of students interacting in the classroom, which includes the interpretation of the teacher and the facts observed. That means it can be written in the immediate moment, as Freeman (1998) noted, or it can be written from memory after classroom events occurred (Lankshear and Knobel, 2004). Through these notes, written data and spoken data are collected as records of the observation. Written data can con-

sist of documents collected for classroom purposes, such as student notes and their work (Lankshear and Knobel, 2004, p. 176). Spoken data can be collected by asking students questions to understand their ideas, values, and attitudes, or by observing conversations occurring in the classroom to identify patterns or how students interact with each other. These three types of data must be coupled with the reflection of the teacher-researcher's own practice (Bortoni-Ricardo, 2008, p. 46) as part of the process of creating knowledge in research. Thus, the research notes can go beyond the descriptive and can contain the teacher-researcher's own interpretations in order to develop a theory about the actions occurring in the classroom. Bogdan and Biklen (2006, p. 120) describe the use of reflections as based on teaching methods, ethical dilemmas and conflicts, and even speculation about themes and patterns to be investigated further in the analysis.

4.4.4 Thematic Analysis

To analyse the data collected through classroom observation, it is necessary to inductively analyse them, having a general sense of the data and coding them, which can be done in themes (Creswell, 2015). For studies focusing on observation in education, one of the approaches is the use of thematic analysis (Cohen et al., 2018). It is a process where themes are identified as patterns, which can be reflected upon to draw conclusions either by themselves or when coupled with other data. Previously in Chapter 3, for the identification and analysis of the studies through synthesis (Subsections 3.4 and 3.5), themes were successfully generated, and led to subthemes that were adopted for the analysis of the data of the SRL. Since these themes were successfully identified in the teaching of MTQA in both TS and NLP, it is appropriate to use the same themes for the thematic analysis as a basis, allowing unique subthemes related to the training course to be identified.

RQs 1 and 4 (**RQ1. - What can be taught to NLP students to encourage them to carry out human evaluation?** and **RQ4 - How effective is the content proposed for HE in an MT evaluation module?**) require the

identification of elements in the training that lead to higher confidence levels and to identify what aspects make the training successful. Thus, much like how the themes were used to respond to SRL-Qs, identified themes and subthemes will be used to respond to RQ1 and RQ4, supplementing the statistical results. Thus, the following codes will be analysed:

a) Teaching Procedures and Methods - TPM - What methodological contributions in the classroom helped students be more confident? What teaching procedures showed evidence of effectiveness?

b) Curriculum and Syllabus - CSL - What particular content from the curriculum led to more effectiveness and willingness to use HE in MTQA?

c) Resources - RE - What resources provided to students led to more effectiveness and willingness to use HE in MTQA?

d) Rationale - RTL - What attitudinal or critical comments made by the students may suggest to willingness or confidence in using HE or metrics in general?

These four themes then lead the thematic analysis. The decision to not adopt Modality (or MOD) was made as modality can be responded to by RE. This particular coding strategy is useful because it can relate directly to how studies in TS and NLP were discussed, leading to an integrated analysis of literature, classroom observation and the statistical results. These themes were applied to the following items, as discussed previously:

a) Observation data: Classroom observation notes on discussions, students' conversations and comments.

b) Written data: Students' notes and assignments

These two types of data will inform the classification of themes into subthemes. Observation data were to be noted immediately. This happened in the classroom when the teacher interacted with students or when the students were completing

activities, with a tablet used for note taking. If notes could not be taken immediately, either when students engaged with the teacher or asked follow-up questions, the teacher typed the notes at the first available opportunity. Written data, on the other hand, were analysed as they were acquired after lessons, either through students sending reports or handing in activity sheets. To organise the data and process them for analysis, the software NVivo was chosen following a recommendation by Cohen et al. (2018).

Chapter 5

Methods - Course Creation - "Development"

5.1 Phase 3 - Development

Based on the previous Analysis phases (1a, Chapter 3, 1b (Section 3.7) and Phase 1c (Section 4.1) and the Design phase (2, Section 4.2), a syllabus and a training course were proposed. This chapter covers both the syllabus and the entire course. To design the syllabus, an IM was adopted, following the elements set in Phase 2. A detailed version can be seen in Appendix F, and a summarised version for better visualisation is shown in Figure 5.1.

The proposed syllabus contained five LOs, which are descriptions of the specific skills and attitudes the student is required to achieve by the end of the training (Diamond, 2008). The LOs for the MTQA module are as follows:

- LO1 - Develop and foster awareness of the role of HE in the development of MT systems.
- LO2 - Develop and foster awareness of the role of AEMs in the development of MT systems.
- LO3 - Develop an understanding of key elements in the development of MT systems, including the purpose of the system, evaluation metrics, type of eval-

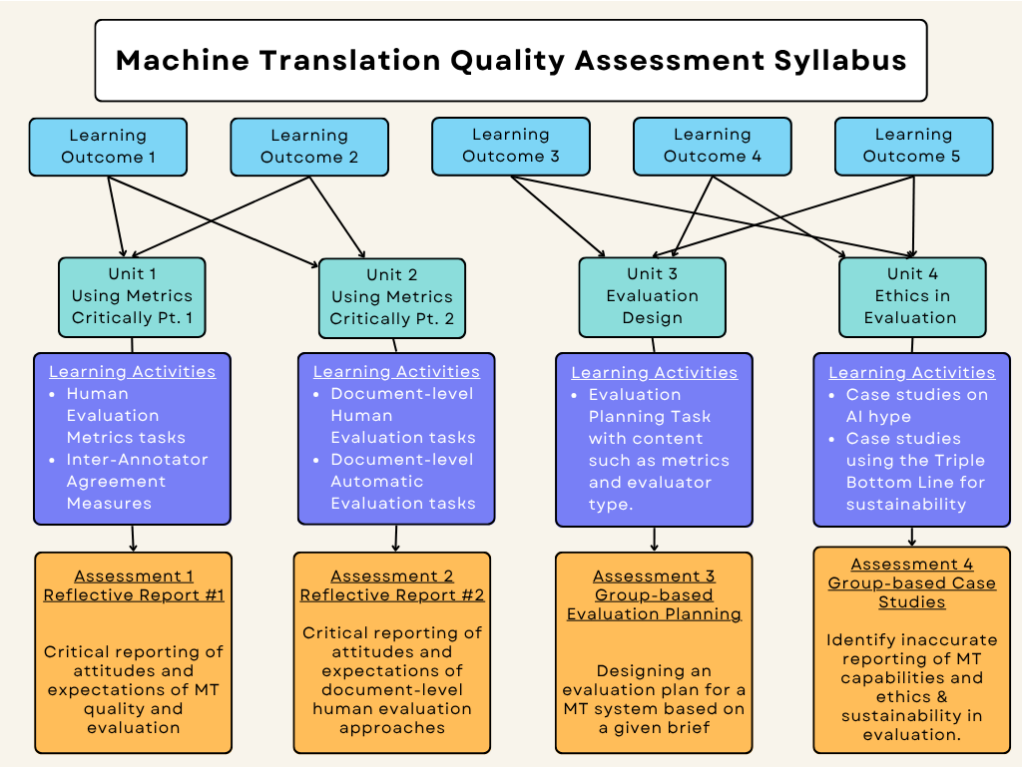


Figure 5.1: Simplified Syllabus containing the learning outcomes, units, learning activities and assessment.

uators, and ethical considerations, to ensure systems are less risky and less biased.

- LO4 - Design replicable evaluation of MT systems, cognisant of the diverse evaluation approaches and types of evaluators in the process.

- LO5 - Report results from the evaluation of an MT system addressing the context, the type of evaluators, and the use case of the MT system.

These LOs originate from the responses related to what constitutes a comprehensive MT evaluation and comments on the importance of incorporating HE in the development of MT systems (see Section 4.1.2). Given the importance of ethical behaviour for developers (Moorkens, 2022), designing training that raises students' awareness of the impact on stakeholders and their roles in the MT development process is crucial for training of this nature. NLP students must be equipped to accurately report what their systems can and cannot do, to avoid hyperbolic reporting and to design rigorous MTQA processes (Toral et al., 2018; Läubli et al., 2020).

I designed four units, as seen in Figure 5.1, each consisting of a 2-hour lesson and a 2-hour laboratory practice, totalling 16 hours of content, following the structure described in the methodology (Section 4.3). The units are:

- Using Metrics Critically Pt. 1 - Lesson 1 (Lecture) and Lesson 2 (Lab) - Section 5.2
- Using Metrics Critically Pt. 2 - Lesson 3 (Lecture) and Lesson 4 (Lab) - Section 5.3
- Evaluation Design - Lesson 5 (Lecture) and Lesson 6 (Lab) - Section 5.4
- Ethics in Evaluation - Lesson 7 (Lecture) and Lesson 8 (Lab) - Section 5.5

The training course was designed to be flexible so that it can be delivered in separate sessions. They are described in the following topics: content (for the lectures), teaching methodology (for the labs), and the assessment. As this chapter presents the planning of lessons, instructions and directions are addressed in writing as "**the teacher**", and the tense alternates between present, future while employing modal verbs to reflect intended instructional guidance and recommendations.

5.2 Unit 1 - Using Metrics Critically Pt.1

In this subsection, the content of the unit (Subsection 5.2.1), the teaching methodology (Section 5.2.2) and the assessment (Subsection 5.2.3) are described, following the outline of Figure 5.2.

5.2.1 Content of Unit 1

Unit 1 presented the following topics. See also Figure 5.2:

(a) Purpose of performing QA (Topics 1, 2 and 3).

(b) Strengths and weaknesses of human evaluation and automatic evaluation (Topics 1,2,3).

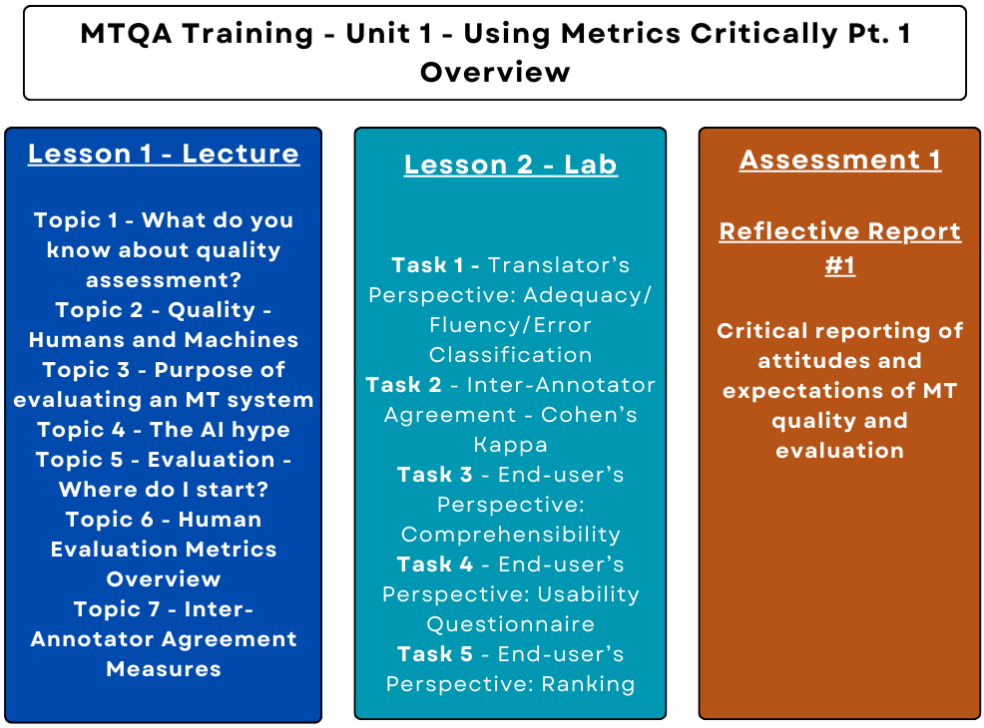


Figure 5.2: Simplified description of topics covered, learning activities and assessment of Unit 1.

(c) AI hype (Topic 4).

(d) Adequacy/Fluency (Topic 6, Task 1).

(e) Error Classification (Topic 6, Task 1).

(f) Comprehensibility (Topic 6, Task 3).

(g) Usability (Topic 6, Task 4).

(h) Ranking (Topic 6, Task 5).

(i) Inter-Annotator Agreement measures (Topic 7, Task 2).

The primary objective of giving an overview of HE metrics is to educate students about why they are implemented in MTQA, how they can be used, and to encourage reflection on their applications. The literature indicates that in TS, translators in training are frequently exposed to AEMs to improve their understanding of how MT systems and MT work (Kenny and Doherty, 2014; Rossi, 2017). Even if MT

developers in training are not directly conducting HE as professionals, perhaps due to lack of multilingual capabilities and linguistic expertise, exposure to HE approaches should make them consider the importance of HE. This exposure aims to challenge their preconceived notions of "quality" and "evaluation" to foster a more adaptable attitude towards applying HE across various contexts and workflows. To follow this assumption, explain, and justify the content choice for the unit, both the literature and the results of the survey with educators will be used.

Defining **(a) the purpose of QA** is important because MT is used in several contexts (Way, 2018). The different contexts can include the development of a new system, a routine check, or a scientific study of the performance or particular use of the MT system (Castilho et al., 2018). While these scenarios require various approaches to MTQA, teaching about **(b) strengths and weaknesses of HE and automatic evaluation** is relevant to the previous topic. As each type of metric has its own use, only the appropriate choices enable valid assessments that can effectively confirm or challenge results of MT systems. It is a particularly useful skill to have, so as not to overestimate the capabilities of systems erroneously (Castilho and Caseli, 2024) through **(c) AI hype**. Furthermore, correct teaching of both types of metrics, human or automatic, helps avoid reliance on one type of metric. In this way, it is possible to prevent the overfocus on AEMs and the devaluation of HE (Moorkens, 2022, p. 126). Finally, understanding how to use metrics can ease the difficulty of implementing different workflows and approaches for the adequate contexts, possible through training (Doherty et al., 2018).

In the survey from Phase 1c, Q29 revealed that both the NLP/CL and TS groups frequently chose **(d) Adequacy/Fluency** and **(i) Inter-Annotator Agreement**. The TS group showed a stronger preference for **(e) Error Analysis** followed by **(f) Comprehensibility**. **(g) Usability** is also a popular choice of content among educators. As through this metric it is possible to see how users interact with the output of MT systems, whether objectively or subjectively, as usability can measure both. Furthermore, there has been an increasing interest in UX (Suojanen et al.,

2014), which represents how UX can contribute as an element of HE.

5.2.2 Teaching Methodology of Unit 1

The purpose of the subsections focused on specific units, in addition to the description of the teaching methodology, is to make recommendations that a teacher should follow. This will be useful in relating how the content is related to the methodological practices, and most importantly it will be useful to use as a basis and compare how practice differed from theory at the implementation (Chapter 6) and evaluation (Chapter 7) of the course. To implement Unit 1, two lessons were planned (Appendix F, Lesson Plan 1 - Lecture and Lesson Plan 2 - Lab). An overview of the structure of lesson 1 (lecture-based) and lesson 2 (lab-based) can be seen prior to its description.

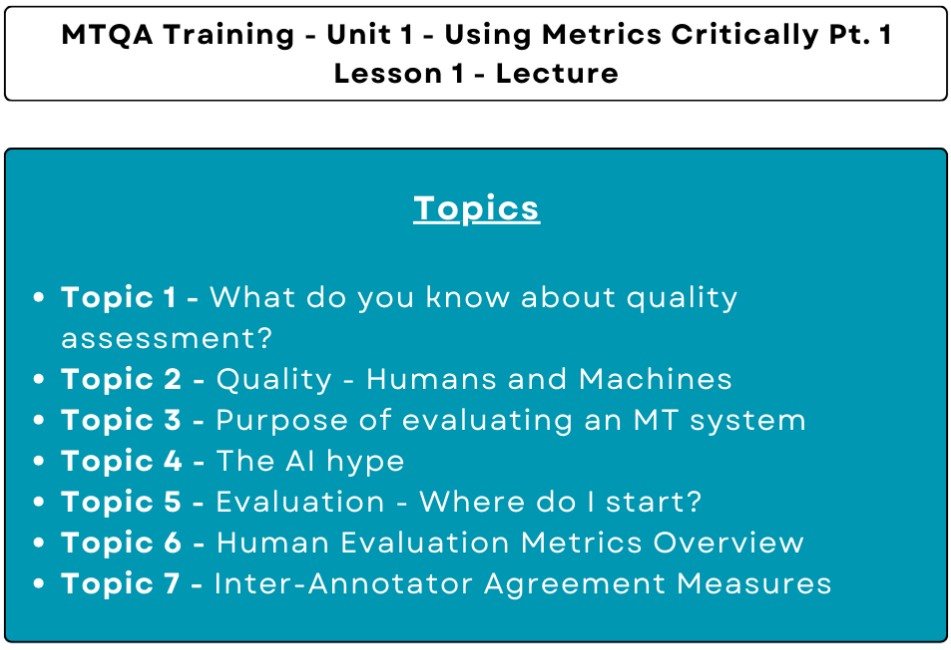


Figure 5.3: Simplified description of topics covered over Lesson 1, delivered as a lecture.

5.2.2.1 Lesson 1 - Lecture

As described in Figure 5.3, the starting topic is about expectations, **Lesson 1** begins with the teacher asking about students' expectations for the training before moving to the measurement students' perceptions aligned with practices from TS (Núñez, 2019; González Pastor, 2021).

The first topic of content to be addressed is to ask what students know about MTQA. It builds on general expectations by screening previous knowledge that students might have had about quality assessment. It is important not to assume that students are empty vessels, as discussed in Subsection 2.2.2 adopting Freire (1996). Following the syllabus, this meant that the lecturer could adjust the depth of the explanations as the lessons go on. For example, it is possible to have NLP students that could have previously acquired knowledge from quality assurance in other NLP tasks.

The lesson introduces more factors that justify the use of HE, such as in-depth analysis of linguistic phenomena and diagnoses of different depths in other respects. It also highlights how evaluation methods should be tailored to the use of the MT system and its end users. For example, if a system is built for professional post-editing or a system built for an end user to gist information, the comprehensiveness of HE must be fit for that purpose.

During the course of the explanation, the teacher has to emphasise that MT systems still rely on humans, requiring different levels of expertise. Since MT systems are not perfect, expectations need to be measured to avoid the idea that MT systems can replace specialised professionals. To sum up this explanation, the flowchart proposed by Doherty et al. (2018) should be presented to the students.

The lesson should continue with an overview of HE metrics, following Castilho et al. (2018). Students should be told that adequacy and fluency are usually measured together, and the teacher must make a quick mention of Likert scales, which eventually become a focus of study over the course of the training sessions. Error classification, acceptability, ranking, and usability should be defined, alongside

appropriate contexts of use. The teacher should explain that the definition of acceptability varies between fields, such as TS or HCI. Likewise, IAA should be introduced, explaining the use of measures such as Cohen's Kappa (Cohen, 1960), Fleiss' Kappa (Fleiss, 1971) and Krippendorf's Alpha (Krippendorff, 1970), by showing the number of evaluators and the scales that are accepted in different communities.

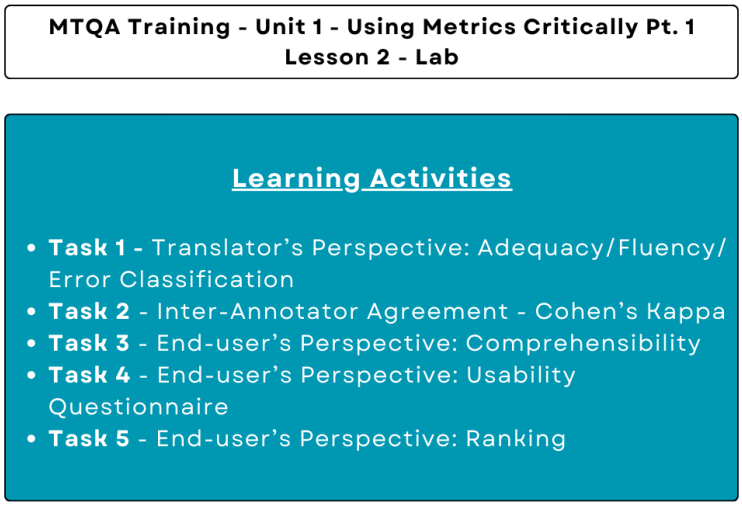


Figure 5.4: Simplified description of tasks covered over Lesson 2, delivered as a lab session.

5.2.2.2 Lesson 2 - Lab

This lesson aims to provide practical exercises to test the theoretical components of MTQA from Lesson 1, and introduce the practice of HE metrics and initial contact with IAA measures (Figure 5.4). The lab contains five tasks: Two tasks focused on testing evaluation metrics from the perspective of a translator; and three tasks focused on testing evaluation metrics from the perspective of an end user. This division of content is intentional, and after each set of tasks, a discussion should be conducted to gather students' attitudes and opinions about the metrics and their expectations for evaluating MT. In addition, some other aspects have to be considered for the use of language pairs in this lab:

i) From a pedagogical (educational) point of view, the recruited students are not expected to be multilingual. The first step of Lesson 1 is for the teacher to ask about

the languages the students speak to fully assert their linguistic profile, but there are other methodological approaches to address this. Since English is the medium of instruction for this training, it is by default one of the languages.

ii) From an operational point of view, it would not be possible to fully evaluate the adequacy/fluency pairing of the metrics with an NLP audience as evaluators. The Tt alone or in conjunction with the ST should be the focus of the task analysis. If the Tt is the sole focus, then, for example, the exercise measures fluency. That also impacts the task related to error classification. As the "mistranslated" or "untranslated" errors can't be detected if the students have no access to the ST.

Task 1 involves an exercise in which students assess segments based on adequacy, fluency, and error classification. They rate each segment on a scale of 1 to 4 for adequacy and fluency and identify types of errors for error classification. The selection of a pair of languages is an important factor for MTQA, as the studies in evaluation design must factor in language pairs to draw conclusions from experiments (Marie et al., 2021). Task 1 is performed on a Google Sheets document, as can be seen in Figure 5.5. The document contains a column for the ST and the Tt, and the remaining columns show three categories: adequacy, fluency, and error classification, as Figure 5.6 illustrates.

Source	MT
É bonita.	It's beautiful.
Mesmo ele mantendo suas costas retas, ele ainda é muito flexível, por isso não atrapalha suas atividades normais.	Even though he keeps his back straight, he is still very flexible, so it does not interfere with his normal activities.
Eles vão te dar um guia de tempo e calor muito geral, mas é	They will give you a very general weather and heat guide, but that's it.
Nós as colocamos de volta por mais 12 minutos, mas observamos que, futuramente, vamos apenas colocar por 24 e	We put them back for another 12 minutes, but note that in the future we will only put them for 24 and hopefully the fryer will tell us when to turn around.
As velas!	The candles!
Elas vão testar sua determinação.	They will test your resolve.
Tudo bem que os sapatos são de salto anabela (eu mesmo prefiro anabela ao estilete, muito mais prático no escalão de saltos), mas eu nunca vi um Dra. em um micro mini saia rosa	It's okay that the shoes are anabolic heels (I myself prefer anabolic heels to stylus, much more practical in the heel range), but I've never seen a Dr. in a micro mini pink shock skirt wearing 12cm shock pink heels before, well,
Usei ela esta noite e a dobradiça quebrou e voou.	I used it tonight and the hinge broke and flew off.
2- Enviar a quebrada de volta, que pela taxa mais barata do	2- Send the broken one back, which for the cheapest USPS rate was $30.68
Esta temporada foi lançada anteriormente como um conjunto de Mídia sob Demanda (MsD) exclusivo da Amazon gravada	This season was previously released as an Amazon-exclusive Media on

Figure 5.5: Interface of Task 1, showing segments of the ST and the TT

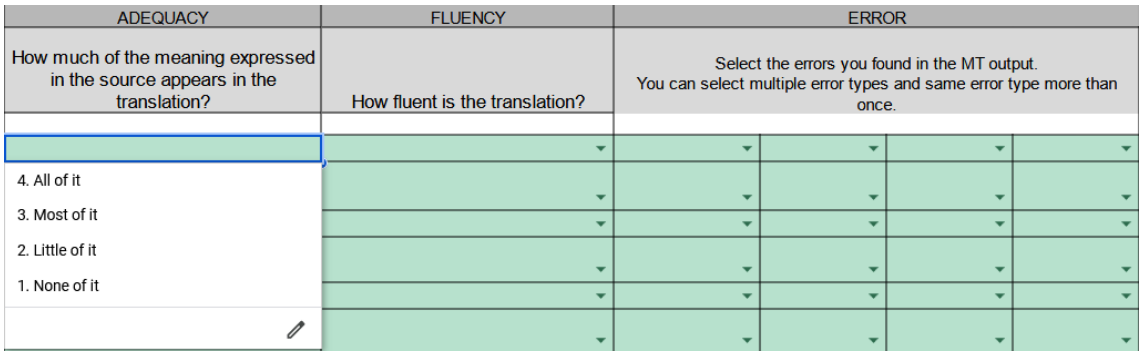


Figure 5.6: Interface of Task 1, showing the categories of adequacy, fluency and error classification with questions to inform the quality assurance and drop-down menus

For the exercise, students are instructed to examine the elements of the UI, including the ST, the MT, the adequacy, the fluency, and the classification of errors. The "source" is in Brazilian Portuguese, and the MT output is in English. With a monolingual English audience in mind, the teacher explains that they only have to perform a fluency assessment because they may not speak Brazilian Portuguese. However, the teacher should state that in a typical MTQA procedure, adequacy is typically included. The teacher also asks students to identify any errors in the MT output if they are able, choosing from "mistranslated", "untranslated", "word order", "word form", "no errors". It is a deliberate choice to provide generic directions, with minor explanations of the Likert scales and no guidelines for classification of errors, as these elements will be introduced and discussed during the course of training as a design and teaching methodology choice.

Task 2 In the following task, students should measure the IAA of their choices, specifically the unweighted Cohen's Kappa using a Google Sheets tab as seen in Figures 5.7 and 5.8:

In this task, the teacher asks the students to sit in pairs and explains that they should retrieve their scores from Task 1. The scores are to be put in columns according to each segment, so a student may assign in the first row the number 3 if they assign that score to that segment. The student's partner would do the same with the score they chose. Consequently, the matrix in Figure 5.8 would update the

T1	T2
3	3
4	4
4	4
4	4
2	1
1	2
4	4
4	4
4	3
4	3

Figure 5.7: Part of the interface of Task 2, showing columns "T1" (translator 1) and "T2" (translator 2) to represent scores drawn from Task 1.

		T2				total
		1	2	3	4	
T1	1	0	1	0	0	1
	2	1	1	1	0	3
	3	0	0	7	0	7
	4	0	0	2	6	8
total		1	2	10	6	19
		1	2	3	4	
	agreement	0	1	7	6	14
	by chance	0.05263157895	0.3157894737	3.684210526	2.210526316	6.263157895
	KAPPA	0.6074380165				
	1	2	3	4		
No:	2	5	17	14		
Per:	5.263157895	13.15789474	44.73684211	36.84210526	100	

Figure 5.8: Interface of Task 2, a matrix illustrating the agreement between two raters (T1 and T2) across four categories, including observed agreement, chance agreement, and a Kappa value. The bottom rows show category-specific totals and proportions.

values. Then, students should discuss the results according to their expectations of agreements and disagreements. For the design of this activity, the scale from Landis and Koch (1977, p. 165) is adopted:

- 0 or negative numbers indicates no agreement.
- 0.01-0.20 as none to slight agreement.
- 0.21-0.40 as fair agreement.
- 0.41-0.60 as moderate agreement.
- 0.61-0.80 as substantial agreement.
- 0.81-1.00 as almost perfect agreement.

The teacher must note that, because the scales are arbitrary, the agreement and their use must be checked within the communities reporting the results. Following the task, a discussion should be prompted with the following questions:

1. "For monolingual evaluations, how challenging was it to evaluate without the text? If you did not perform monolingual evaluations, how did the overall difficulty compare?"

2. "Is there a correlation between adequacy/fluency scores and the number of errors identified? Specifically, did lower scores in adequacy/fluency correspond with more errors being identified, and did higher scores correspond with fewer errors?"

3. "How would a deep understanding of both languages impact the evaluation? What would happen if someone without any skills attempted to assess the quality?"

4. "When calculating IAA, what discrepancies (if any) were noticed in the assessments? What do you think caused them?"

These questions help discussing Tasks 1 and 2. It is pertinent to consider these questions from the perspective of an expert evaluator. The teacher should tell students that translators would give detailed insights as they can describe the linguistic phenomena they encounter to correlate certain factors, such as the number of errors to adequacy/fluency scores or reasons for agreement and disagreement. In this discussion, the teacher may reveal to the students the MT systems used in the design of the task, to start a discussion about biases. To do so, the teacher may reveal their version of the sheet, which identifies the MT systems, as Figure 5.9 shows.

Task 3 From Task 3 on, the focus is changed to the perspective of a non-expert end user of MT. In this task, the students are presented with a document, in one tab of which three columns may be seen, as indicated in Figure 5.10.

Eu assisti novamente este filme uma e outra vez porque Naruto e Hinata são meu par favorito.	I watched this movie again and again because Naruto and Hinata are my favorite couple.
O enorme órgão musical preenche o espaço e adiciona romance!	The huge musical organ fills the space and adds romance!
E a partir de hoje, estou convertida.	And as of today, I'm converted.
	MT1 - General - Google MT
	MT2 - General - Bing MT
	MT3 - Specific domain - eTranslation

Figure 5.9: Interface of the teacher's version of the document sheet of Task 1, with colours identifying the MT systems.

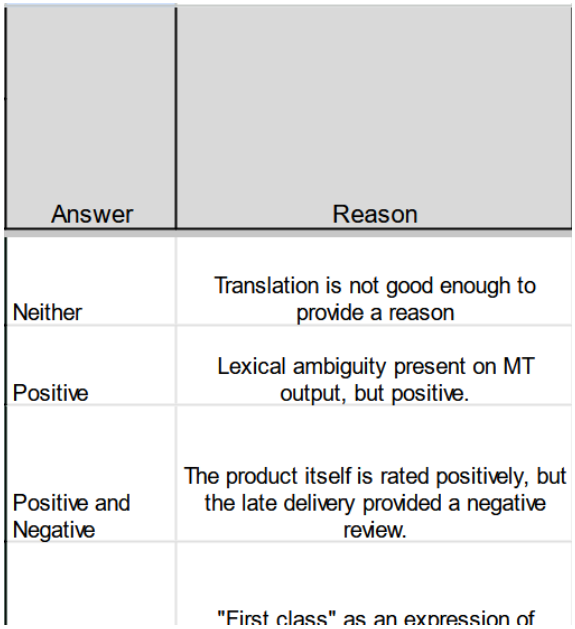
Source	MT	Task You will read reviews of products or services. What type of description or judgment can you detect reading the review?
Hem estat agust i menjat molt bé.	We were August and ate very well.	
La regalé a mi pareja y la llevamos a todos los viajes, es muy ligera y fácil de usar, y las fotos salen preciosas!	I gave it to my partner and we took it on every trip, it's very light and easy to use, and the photos are beautiful!	A positive judgment A negative judgment Both Neither positive, neither negative
Tá an Guinness Íontach ar fhád le neart daoine aitiúil istigh ag ól.	The Great Guinness is on the way with plenty of locals drinking it.	
os lanches são maravilhosos (recomendo especialmente o ciabatta de tofu, a esfiha de PTS e a sopa paraguaia);	the snacks are wonderful (I especially recommend the tofu ciabatta, the PTS esfiha and the Paraguayan soup);	

Figure 5.10: Interface of Task 3, where users assess the comprehensibility of MT output by categorising multilingual reviews based on their sentiment or descriptive nature, where evaluators can choose from positive, negative or neutral judgements from a drop-down menu.

From the perspective of an end user, the design of this task in terms of language pair is much easier, as the focus remains on the MT output. Because of this freedom, segments from languages other than Brazilian Portuguese were considered when designing the task. Additional languages such as Italian and Spanish were added. In addition, to trigger a discussion of quality expectations of MT, low-resource languages of different degrees were selected. For low-resourced languages such as Irish (Costa-jussà et al., 2022) and Catalan (Melero et al., 2023) NMT systems' performance depends on large datasets, so it is essential to demonstrate how the quality drops when resources are limited (Ranathunga et al., 2023). To collect the pedagogical dataset for this task, selected reviews from Google Maps and Amazon were used, so the MT systems match where they were collected.

Thus, in this task, students examine sentences taken from product reviews and

focus solely on the MT output to answer the question in the third column: "You will read reviews of products or services. What type of description or judgement can you detect reading the review?" The students must choose between: positive; negative; both positive and negative; and neither positive nor negative. The teacher should have their own version with possible answers, to guide the students on the questions, as Figure 5.11 shows.



Answer	Reason
Neither	Translation is not good enough to provide a reason
Positive	Lexical ambiguity present on MT output, but positive.
Positive and Negative	The product itself is rated positively, but the late delivery provided a negative review.
	"First class" as an expression of

Figure 5.11: Interface of Task 3 for the teacher only, reference columns for instructors to guide students' questions during the comprehensibility evaluation task. It provides the correct answers and corresponding reasons, enabling teachers to clarify doubts and support students in understanding the quality and clarity of MT.

Task 4 - Subsequently, Task 4 is performed in relation to Task 3. In this task, students must respond to a usability post-task questionnaire. They must have reflected on their experience of the MT systems used in Task 3, rating statements on a scale of 1 to 5, as can be seen in Figure 5.12.

Task 5 - The last metric to be tested by the students is ranking, as Task 5. Ranking is typically used as the ranking in relation to a specific aspect or metric, such as ranking according to adequacy or fluency (Castilho et al., 2018). For this exercise, the use of ranking is based on no specific criteria. It shows students that

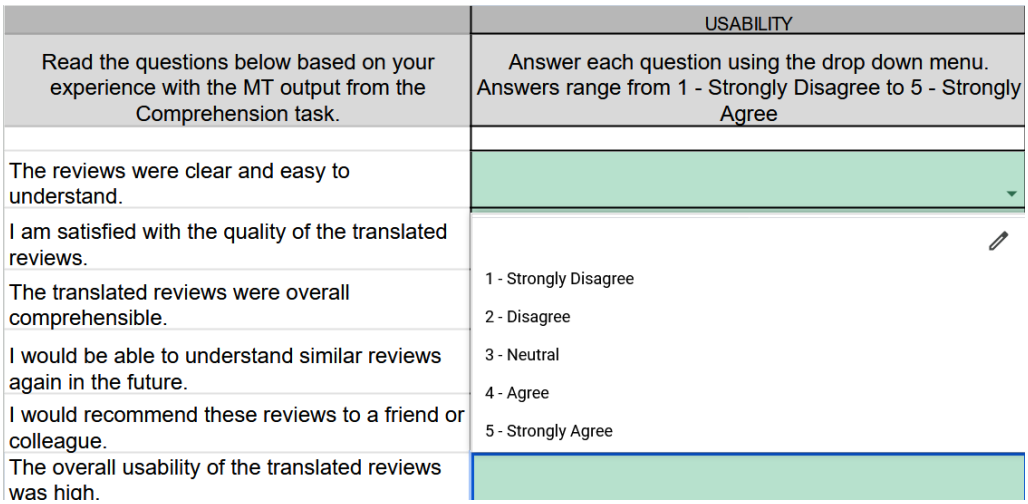


Figure 5.12: Interface of Task 4. It shows a Likert scale questionnaire where users rate their experience with the MT output from the comprehension task. Questions assess clarity, satisfaction, overall comprehensibility, potential for future understanding, recommendation likelihood, and the overall usability of the translated reviews.

if there are no specific linguistic criteria attached to it, if the systems being ranked are safe and acceptable to use, it becomes a matter of personal preference. Students are given columns with Source, MT1, MT2, and "Best translation", where students choose either MT1, MT2 or both, as can be seen in Figure 5.13

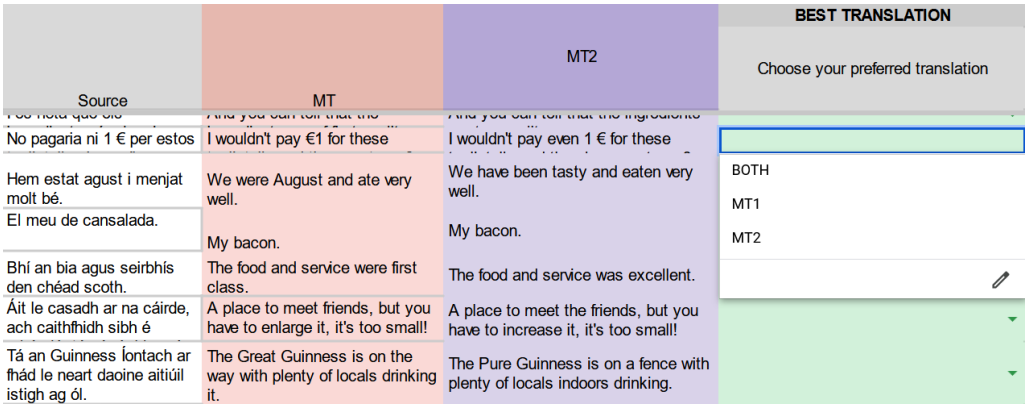


Figure 5.13: Interface of Task 5. Students evaluate and select their preferred translation or indicate if both are equally good. The task focuses on assessing MT outputs from an end user perspective, with no specific criteria given.

Following this task, the teacher starts a discussion again, but with different questions to understand how students evaluated from the perspective of an end user, such as:

1. "How challenging was comprehensibility? Considering different languages were

evaluated, do you think the language affected the MT?"

2. "What could be the potential consequences for an end user if a translation inaccurately conveys the sentiment of a review?"

3. "How could this affect an end user's purchasing decision or perception of a business? What risks are associated with poor translation quality in this context?"

4. "Regarding usability, what role does ease of understanding and use play in the perceived reliability of the information?"

5. "How does the concept of usability in translation differ when considered from an end user's perspective versus a translator's perspective? How would usability be tested for a translator?"

6. "How was it to rank translations without clear criteria? Would it be better to implement criteria?"

The discussion that stems from these questions not only allows students to discuss what they assessed as an end user, even drawing from their own experiences with MT, but it also makes them draw parallels to evaluation from the perspective of a translator. The questions revisit important topics from previous tasks, allowing students to reflect on these related topics, such as expectations of low-resource MT systems or subjective criteria applied to certain metrics.

5.2.3 Assessment of Unit 1

The assessment of this unit should be performed after Lesson 2 (the lab session) and is included in the lesson itself. The assessment should be carried out in the classroom, and consists of a short reflective written document of 200 to 500 words. It is inspired by a practice normally adopted in translation pedagogy, in which students are tasked with writing a reflective report about their experiences after participating

in laboratory exercises. The conclusions are drawn from Doherty and Kenny (2014) and Moorkens's (2018) findings, particularly with regard to the use of assignments that encourage reflection. These assignments help to better understand students' experiences and evaluate the module. Furthermore, as highlighted in Rossi (2017) and Moorkens (2018), the attitudes and feelings of students toward the technologies involved in the assignments provide two valuable types of insight: The first are pedagogical insights that can be used to adapt teaching strategies, explanations, and materials. The second type are insights about the subject matter content needed to better understand the relation of the student and the expertise they are learning with the materials taught.

Unlike previous studies that concentrated on translation students' apprehensions about technology use, this unit proposes that NLP students should think about how their creations impact those who use them. Moreover, following Moorkens (2018), encouraging students to reflect on the quality expectations of the MT systems allows students to better understand how to use technology, in this case, to better understand how to create technology. In this way, the students would examine it from a viewpoint distinct from that of the translators. In this training, they should learn not to underestimate MT as a task, as it remains an unsolved problem in NLP, as evidenced by Moorkens et al. (2024b, p. 18). With these studies and with the intention of making students reflect on these issues, the following questions are suggested.

1. What were your expectations in terms of translation quality before doing this exercise?

2. How did you adapt your evaluation approach when considering the translator's perspective versus the end user's perspective?

3. Did you expect the IAA results you have obtained? Comment on the differences between your scores and of your colleague.

4. Will this change how you use MT?

These questions can promote reflection on quality assessment in different ways.

Question 1 is intended to examine in particular how NLP students may feel toward MT, as it is widely used for different use cases (Way, 2018), such as grasping the rough idea of something by gisting (Bowker and Ciro, 2019). This brings clarity to the end user's perspective, to understand the needs of people who use and benefit from MT systems (Nurminen and Koponen, 2020).

Question 2 emphasises the different roles with different tasks, since the objective was to examine both the expertise required for an informed evaluation and how lack of expertise can limit the in-depth understanding of certain phenomena.

Question 3 aims to examine the subjective nature of HE (Castilho et al., 2018) with which people may agree or disagree. This question was designed to remind NLP students that subjectivity is indeed challenging but that only an expert would be able to explain their scores according to their linguistic and translation skills.

Question 4 in a sense allows the NLP students to express their personal tastes when using MT. From a pedagogical point of view as the researcher who is the teacher, this question was devised to get initial qualitative evidence from their responses. That is, evidence of a change in attitudes or effectiveness of the training from the first lesson.

Questions 1 and 3 are useful for examining students at the beginning of the training. While Questions 2 and 4 provide more direct information on whether the training has effected change in the students by adding layers to MTQA, such as who is evaluating and whether the training has provided any tool or approach that might influence their use of MT and MTQA. These questions are useful for the teacher. The outcome of these questions is described in Chapter 7 (Sections 7.2 and 7.3), which comprehensively organises the results of the assessments and how the training influences attitudes and confidence of students undertaking the course described in this chapter.

5.3 Unit 2 - Using Metrics Critically Pt.2

This subsection presents the content of Unit 2 (Section 5.3.1, the teaching methodology (Section 5.3.2), and the assessment (Section 5.3.3), following the structure seen in Figure 5.14.

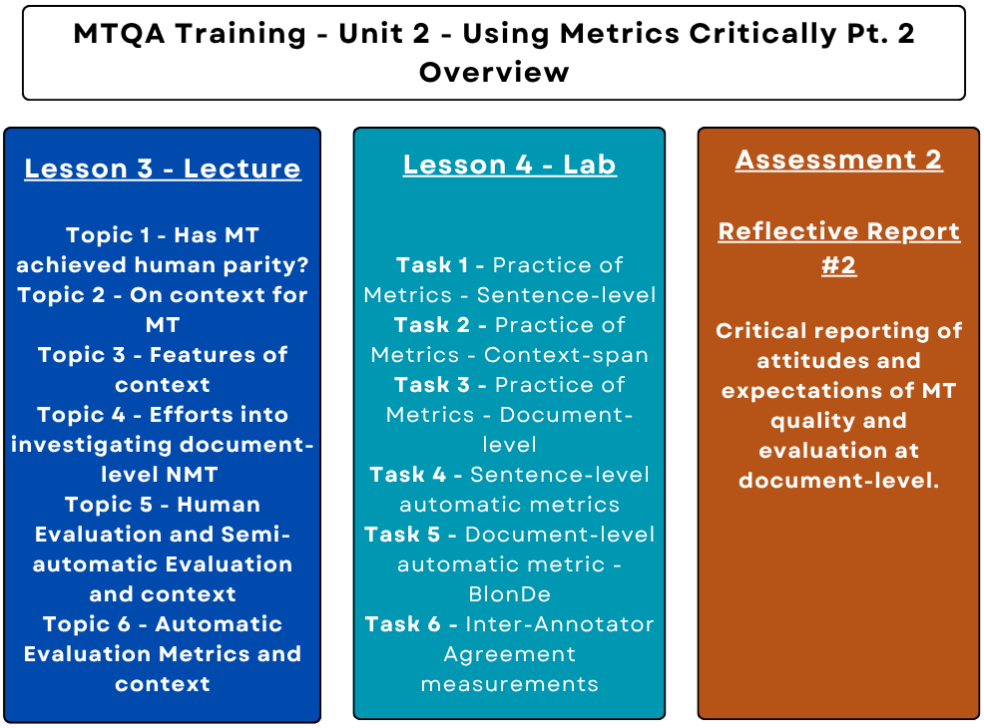


Figure 5.14: Simplified description of topics covered, learning activities and assessment of Unit 2.

5.3.1 Content of Unit 2

(a) Human Parity Claims in MT (Topic 1).

(b) Limitations of Sentence-Level in MT (Topic 2 and 6, Task 4 and 5).

(c) Dimensions of Context in MT (Topic 2 and 3).

(d) Document-Level research in NMT (Topic 4).

(e) Contextual Factors in Human and Semi-Automatic Evaluation (Topic 5, Task 1, 2, 3 and 6).

(f) Document-level Automatic Evaluation Metrics (Topic 6, Task 5).

Castilho and Knowles (2024) explain that much of the research in the MT evaluation field remains focused on the sentence level and that a broader evaluation requires more attention. This is confirmed by Phase 1c (Section 4.1.2), which revealed that TS teachers consider interest in context within MTQA one of the most probable trends, while NLP teachers perceived context in MTQA to be equally important as adding UX as a factor in evaluation. After all, as Castilho and Knowles (2024, p. 3) explain, it is not known whether document-level NMT models are able to rely on context to solve ambiguities or whether they still need additional improvements.

The starting point of the unit is to investigate **(a) Human Parity Claims in Machine Translation**, stemming from studies such as Hassan et al. (2018). It is a great problem for shared tasks when strong claims are made about the capabilities of MT systems when they are detached from a rigorous and comprehensive evaluation, particularly when they are based on AEMs and lack HE (Moorkens, 2022). This discussion leads to the identification of the **(b) Limitations of sentence-level evaluation in Machine Translation**. To show the limitations of sentence-level MT to students, it is useful to identify how the concept of "context" as developed in TS. In TS, the notion of "context" has been informed by perspectives from several other disciplines (House, 2014), impacting MT research. Castilho and Knowles (2024) suggest the typology offered by Melby and Foster (2010) is useful in identifying the types of context that have been explored in MT research, leading to **(c) Dimensions of Context in Machine Translation**, which are defined as intersentential context (Castilho and Knowles, 2024, p. 3). That is, when a main sentence is provided together with its surrounding context, with preceding and following sentences from the same document. In the context of NLP techniques for MT, it would range from the unit of translation being a paragraph or a whole document, or extending the length of tokens (Castilho and Knowles, 2024, p. 4-5). Thus, on that basis Castilho and Knowles (2024) listed a considerable number of studies discussing **(d) Document-Level research in Neural Machine Translation**.

In this topic, surveys from the literature are discussed, listing even more specific aspects of discourse that have been implemented in document-level NMT models.

These four topics provide an overview that demonstrates to students the progress and limitations of document-level implementations of NMT models. With that foundation provided, more attention can be given to HE methods and AEMs. That is critical because document-level HE approaches and AEMs are included in the lab in some manner. Because assessing the efficiency of these approaches is still a challenge (Castilho and Knowles, 2024, p. 10), such an overview and some degree of practice would inspire students to explore this evolving area of MT development and assessment. In the case of **(e) Contextual Factors in Human and Semi-Automatic Evaluation**, the conclusions of the work of Castilho et al. (2020) are essential to study and implement the "context span" in the HE using metrics such as adequacy and fluency, which the students would have learned in the previous lesson.

5.3.2 Teaching Methodology of Unit 2

Consequently, to implement the content described in Section 5.3.1, two sessions were designed (Appendix F), with Lesson Plan 3 - Lecture and Lesson Plan 4 - Lab. The overview for the lecture, **Lesson 3**, is described in Figure 5.15.

5.3.2.1 Lesson 3 - Lecture

The first topic of Lesson 3 to be discussed involves reflection on the current state of MT. The teacher asks the students if they are familiar with the concept of "human parity". Using slides, the teacher reviews studies that made claims that MT has achieved human parity and studies that contested those claims. From this topic, the teacher should ask the students about their experience in the lab session of Lesson 2. Although NLP students might not have the ability to fully assess translations with precise linguistic nuance, by checking fluency, they can confirm suspicions that MT is not perfect.

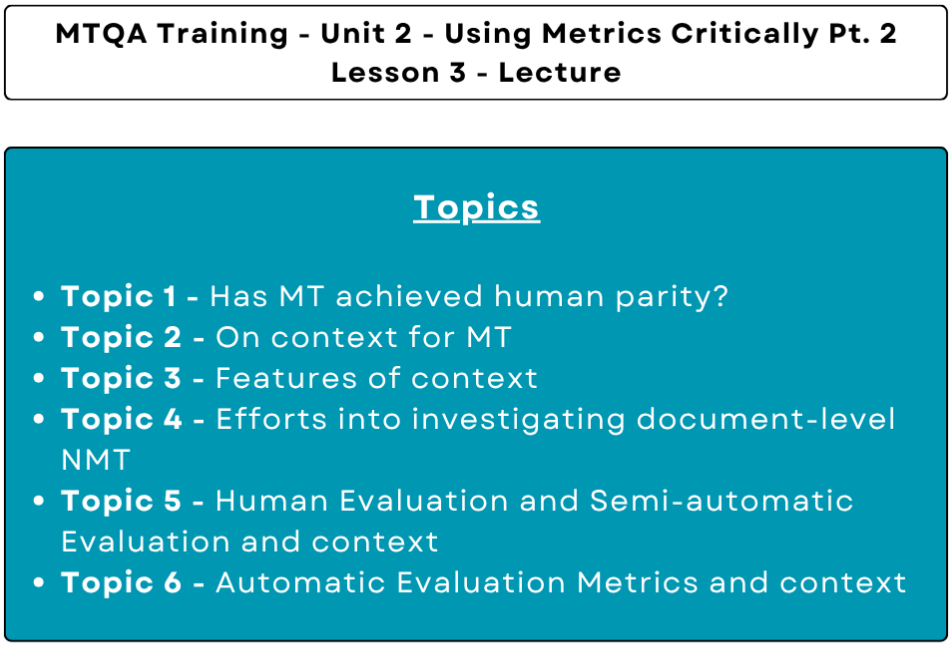


Figure 5.15: Simplified description of topics covered over Lesson 3, delivered as a lecture.

This is a perfect hook for the teacher to discuss NMT systems. The neural paradigm of MT systems had been mostly implemented at sentence level (Castilho and Knowles, 2024), an aspect that students likely would notice from the lab activities of Lesson 2, because the activities involved the evaluation of sentences without any context. That allows the teacher to discuss what technical approaches are used to insert context in MT systems. For example, works that apply a context span to limited sentence pairs (Tiedemann and Scherrer, 2017; Bawden et al., 2018; Müller et al., 2018), the use of documents substructures (Dobreva et al., 2020), or attempts to go to a specific span (Junczys-Dowmunt, 2019; Voita et al., 2019).

After reporting the studies mentioned previously, the teacher should investigate the concept of context itself. This is how the concepts provided by Melby and Foster (2010) are relevant. Co-text, chron-text, rel-text, bi-text, and non-text must be explained. Afterwards, these concepts should be applied to different instances in MT development. For example, following Castilho and Knowles (2024), the concepts of intersentential context, world knowledge, and terminology are highlighted as to how

these context concepts appear in practice in MT research. Knowing these concepts, the teacher should be able to describe how some of these concepts are applied in MT research through surveys that were investigated in depth by Popescu-Belis (2019), Kim et al. (2019), Lopes et al. (2020), Maruf et al. (2021), Jin et al. (2023) and Castilho and Knowles (2024), the latter being the basis of the lesson.

With such an overview completed, the teacher can finally focus on evaluation approaches, both human and automatic. For this lesson, the topics of evaluation methodologies, the amount of context needed for in-depth linguistic analysis, and how to use test suites are presented. The studies of Läubli et al. (2018); Toral et al. (2018); Graham et al. (2019); Läubli et al. (2020) and Gilbert (2023) must be mentioned, but will only receive more time to actively discuss the following unit, evaluation design (available in Section 5.4). These studies suggest approaches that go beyond the sentence level, which is how the lecture showcases the available approaches for HE, for example.

For HE approaches, the teacher starts the explanation with the study by Castilho et al. (2020). Using this study, the teacher should be able to suggest a term adopted in the paper to refer to the context: context span. According to the results, 33% of sentences' context errors could be resolved with two preceding sentences and other sentences required further context to resolve the ambiguities is essential and must be clarified for students. Castilho (2020) and Castilho (2021) should be mentioned for their robust design and results. The teacher should also explain the scales used in HE, employing Direct Assessment (DA) (Graham et al., 2013). This type of scale has been applied to shared tasks, particularly WMT 2019 (Junczys-Dowmunt, 2019), providing more context for students about their role in shared tasks.

Test suites must also be mentioned as a way in which context-related issues can be found in MT output. These challenging test sets, although limited in availability (Castilho and Knowles, 2024), allow the identification of specific language issues at the discourse level that MT systems are not yet able to detect. For example, insights from Rysova et al. (2019) should be shared, as the authors disclose that "qualitative

analyses of document-level errors in MT outputs" are to date quite rare (Rysova et al., 2019, p. 1), further reinforcing test suites as a potentially useful resource. Castilho et al. (2021) may be used to explain the data collection for the test suite, highlighting that they were from available sources and that it took a considerable amount of work to compile such a resource. The contributions of Castilho et al. (2021) should be explained in the lesson with the purpose of explaining the issues found in the corpus, the number of sentences, the annotation process, and the process of verifying agreement.

The final content to be covered shifts to AEMs. Thus, very briefly BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), chrF (Popović, 2015), TER (Snover et al., 2006), and BERTScore (Zhang et al., 2019) can be mentioned. Their integration at the document level should be discussed as well (Wong and Kit, 2012; Gong et al., 2015; Xiong et al., 2019; Liu et al., 2020; Saunders et al., 2020; Vernikos et al., 2022). In contrast, there are metrics that concentrate solely on the document level, such as BlonDe (Jiang et al., 2022). BlonDe measures similarity with reference-based metrics with respect to entity, tense, pronoun, and discourse markers. The teacher should mention that, despite BlonDe showing lower performance in comparison to the previous metrics, the other metrics still score sentences one sentence at a time, rather than the entire document (Castilho and Knowles, 2024, p. 14). DiscoScore allows customisation for target nouns or entities (Zhao et al., 2023) in two variants, DS-Focus tracks how coherent the entities or nouns are across sentences, while DS-Sent analyses the relationship between sentences through a graph. Finally, SLIDE, proposed by Raunak et al. (2024) should be presented. It stands for Sliding Document Evaluator and works by defining a context window size to be calculated using COMET with a single input, showing promising results (Castilho and Knowles, 2024, p.15). The teacher should mention that, just because a single metric targets different levels of discourse, evaluating at the document level does not automatically make a metric better. The teacher should stress that and a combination of metrics is ideal and should be chosen to fit the scenario.

5.3.2.2 Lesson 4 - Lab

The lab session aims to expand the practical exercises from Lesson 1, adopting the methodologies presented in Lesson 3. It also aims to introduce the practice of using AEMs, with more attention devoted to a document-level AEM, BlonDe. In total, there are six tasks, which can be observed in Figure 5.14. This lesson uses most of the pedagogical dataset described in section 4.2.3.

Task 1 - The first task, along with Tasks 2 and 3, occurs in the platform Qualtrics. Task 1 consists of 11 sentences extracted from different domains, such as literature, user-generated content, and scientific texts. Task 1 contains instructions that show the evaluation instructions, and a box with a column "Source" for the ST and a column with MT output, as seen in Figure 5.16.

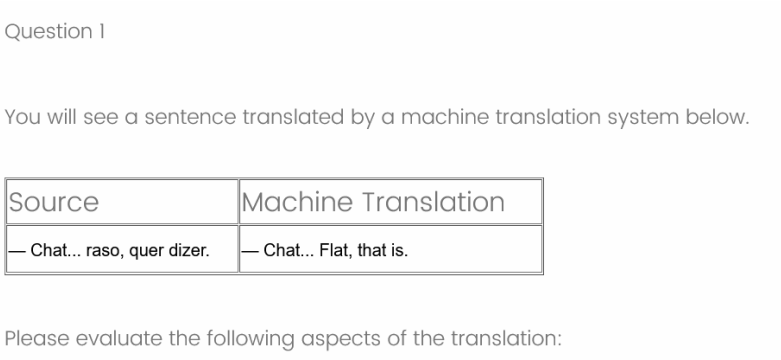


Figure 5.16: Interface of Qualtrics with explanations related to Task 1

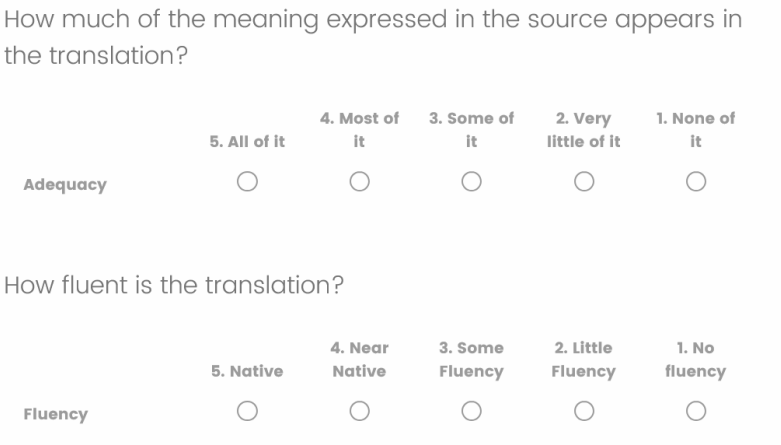


Figure 5.17: Interface of Qualtrics for the evaluation of adequacy and fluency

Each question contains one sub question about fluency and one about adequacy, to be evaluated on a scale of 1 to 5 as per Figure 5.17. In other tasks, the scales differ, so students can examine how they affect their evaluation, in order to experiment how settings can affect their experience based on the conclusions of the study conducted by Knowles and Lo (2024). In this case, the intent is to allow for a neutral option with the middle point. In total, there are 13 questions for the students to answer, each with a sub question related to adequacy, fluency, and a sub question about their confidence in evaluating the MT sentence, as seen in Figure 5.18.

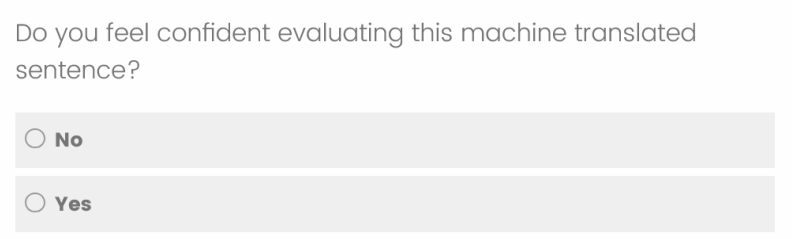


Figure 5.18: Question on Qualtrics about confidence when evaluating the sentence

After all sentences are evaluated in the first set of questions and sub questions, the teacher should have a discussion. The topics are: i) how students feel about different platforms being used (Google Sheets used in Lesson 2 and Qualtrics used in Lesson 4); ii) the implications of having a middle option in the Likert scale and; iii) whether they are able to understand the intended meaning and to identify the types of issues found in the sentences.

Task 2 - The second task follows a similar format but modifies the approach for the evaluation, adding context span. Figure 5.19 shows the instructions including one preceding and one succeeding sentence to add more context to the evaluation.

The sentences used are the same sentences from the previous activity. Since the focus is on the methodology, using this same text allows students to perceive a sentence differently when more context is added. Although the experiments conducted by Castilho et al. (2020) demonstrated that two sentences (either preceding or succeeding) are enough for most sentences to be correctly evaluated, the sentences vary so that students can draw their own conclusions. The discussion after the task will

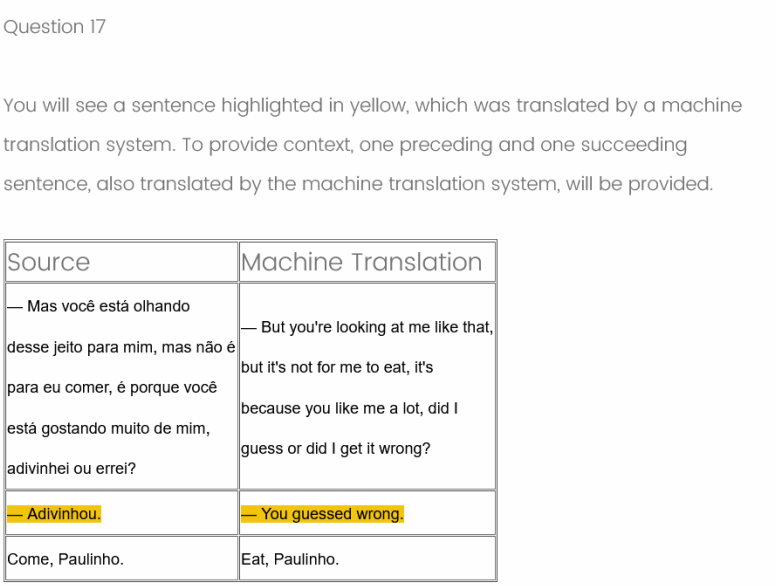


Figure 5.19: Instructions on how to evaluate a MT sentence highlighted in yellow used the context span approach

include the questions: i) how useful is the amount of context span length; ii) what types of issues found are context related, such as terminology or lexical ambiguity; iii) Since the Likert scale of Task 2 has no neutral point (1-4), how does it compare to a scale that has a neutral point (1-5).

Task 3 - The third and final task on HE with additional context span consists of the evaluation of sentences with the full text available. In this task, the students also have to provide a score for the entire document. First, students should be shown a description of the task, as seen in Figure 5.20:

In sequence, students should examine the entire text until they find sentences numbered and highlighted in yellow, as Figure 5.21 shows:

For the rating of adequacy and fluency, students have to use a DA scale from 1 to 100, as may be seen in Figure 5.22.

The DA scale is implemented in the task as it was progressively implemented (Bojar et al., 2016, 2017) in the shared WMT task up to the point where it included full document evaluation (Barrault et al., 2019). As such, it is a relevant methodological approach to introduce to students, shown in their UI as Figure 5.23 regarding adequacy, followed by a fluency question using the same methodology.

Question 29

You will highlighted sentences in yellow, which was translated by a machine translation system. To provide context, the whole document, also translated by the machine translation system, will be provided.

Source	Machine Translation
Come, Meu Filho - Clarice Lispector	Eat, My Son - Clarice Lispector
O mundo parece chato mas eu sei que não é.	The world seems boring, but I know it's not.
Sabe por que parece chato?	Do you know why it seems boring?
Porque, sempre que a gente olha, o céu está em cima, nunca está embaixo, nunca está de lado.	Because every time we look, the sky is above, it's never below, it's never to one side.

Figure 5.20: Question on Qualtrics describing the task of a document-level evaluation activity

Quando dizem para ele, é só acreditar, pra ele nada precisa parecer.	When they tell him, he just has to believe them, nothing has to seem that way to him.
Você prefere prato fundo ou prato chato, mamãe?	Do you prefer deep dish or flat dish, Mum?
1. — Chat... raso, quer dizer.	1. — Chat... Flat, that is.
— Eu também.	— Me too.
No fundo, parece que cabe mais, mas é só para o fundo, no chato cabe para os lados e a gente vê logo tudo o que tem.	In the deep one, it looks like it fits more, but it's only towards the bottom, in the flat one it fits to the sides and you can see everything there.

Figure 5.21: Example of a numbered sentence highlighted in yellow for students to evaluate

Moreover, as the students evaluate different sets of questions, the scales are modified with different settings. This is designed to make students realise that these settings affect the evaluation process well. For example, Figure 5.24 shows a modified scale with more increments (0, 25, 50, 75 and 100) and the exact number the user is choosing to select.

The settings for each text are:

- Text 1 - One increment (100), two labels (For adequacy, “None” and “All of it”, For fluency, “No Fluency” and “Native”, no custom position, no value shown to evaluators.
- Text 2 - Four increments (25, 50, 75, 100), two labels (For adequacy , “None” and “All of it”, For fluency, “No Fluency” and “Native”, no custom position,

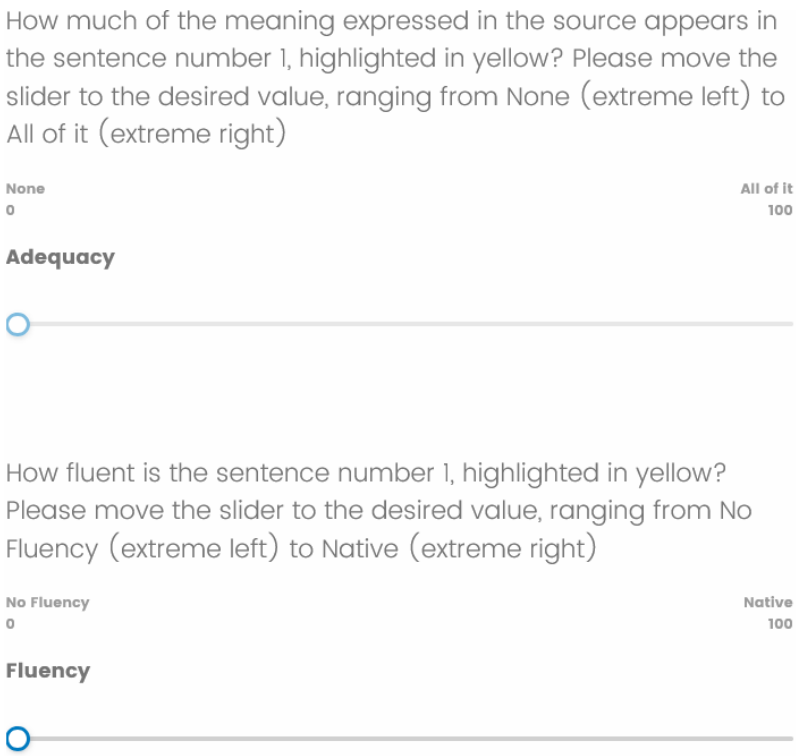


Figure 5.22: DA scale presented to students for rating adequacy and fluency

show value to evaluators.

- Text 3 - Ten increments (10, 20, 30, 40, 50, 60, 70, 80, 90, 100), two labels (For adequacy, “None” and “All of it”, for fluency, “No Fluency” and “Native”, no custom position, no value shown to evaluators.

- Text 4 - One increment (100), five labels (For adequacy, “None”, “Little of it”, “Some of it”, “Most of it”, “All of it”, for fluency, “No Fluency”, “Little Fluency”, “Some Fluency”, “Near Native”, “Native”), no custom position, no value shown to evaluators.

- Text 5 - four increments (25, 50, 75, 100), two labels (For adequacy, “None” and “All of it”, for fluency, “No Fluency” and “Native”, custom position at 50, show value to evaluators.

- Text 6 - four increments (25, 50, 75, 100), two labels (For adequacy, “None” and “All of it”, for fluency, “No Fluency” and “Native”, custom position at 50, no value shown to evaluators, scale snaps to increments (e.g., if evaluator

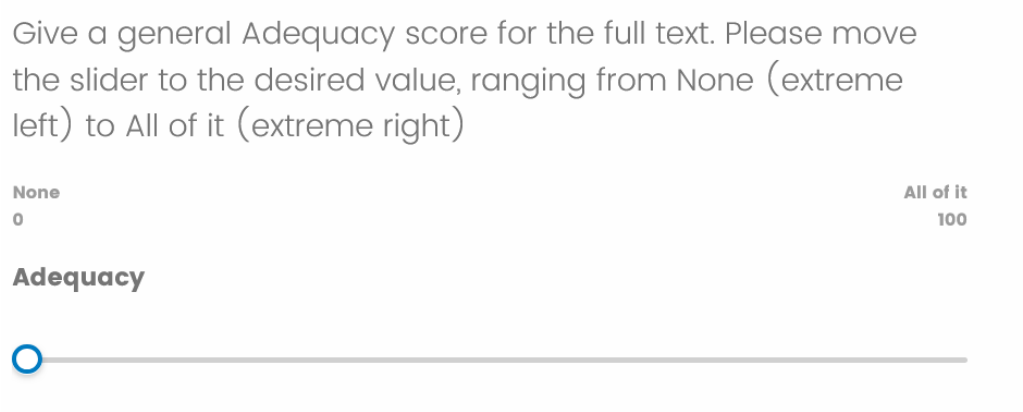


Figure 5.23: DA scale to rate the full document in respect of adequacy.

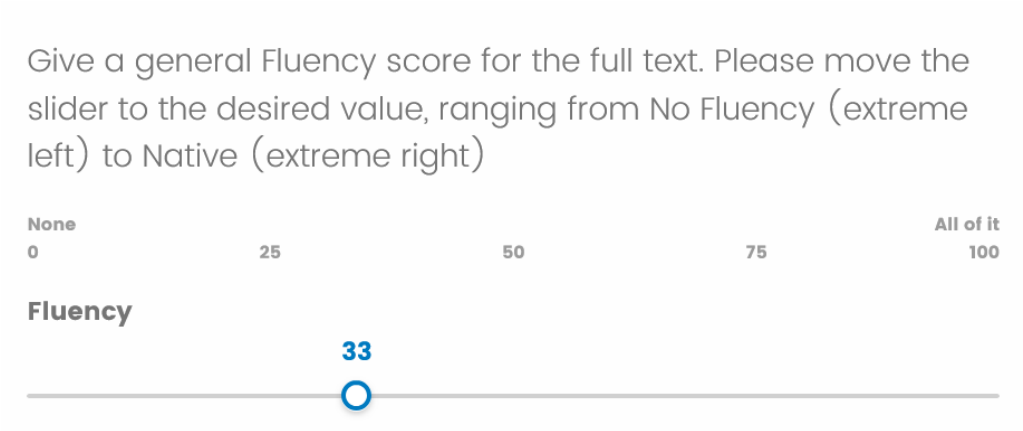


Figure 5.24: DA scale showing four increments (0, 25, 50, 75, 100), two labels ("None" and 'All of It"), no custom position with scaling showing the selection upon clicking and a value shown to evaluators (in this case, "33").

chooses 17 in the scale, scale snaps to 25; if evaluator chooses 72, scale snaps to 75; only 0, 25, 50, 75 and 100 can be chosen).

After testing these different settings with a full document DA, the teacher should organise a discussion. The teacher should ask students about: i) differences between each customisation of scales; ii) potential challenges the scales might bring (such as using phones or not having instructions); ii) the comparison of evaluating sentences with the evaluation of the whole document; iv) if any symptoms of fatigue, boredom, or any other aspect that may have impacted the evaluation are present.

Task 4 - Concluding the HE element of the class, the tasks would focus next on AEMs. Using datasets (Section 4.2.3), students have access to the source, target, and reference texts. They should be used on the website MATEO (Vanroy et al., 2023).

The choice to use MATEO is a result of adopting Principle 3 of the SRL conducted in the thesis (Section 3.6), in which good evaluation platforms offer good UX, such as the accessible scoring of AEMs. As such, for this task, students examine the scores calculated for each AEM: BERTScore (Zhang et al., 2019), BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020), ChrF (Popović, 2015) and TER (Snover et al., 2006).

The purpose of this task is for students to examine the features of AEMs, strengths, and limitations. Students have to compare the metrics using the features shown in the website: i) bar plots; ii) table indicating scores with confidence intervals and p-values added; iii) the metric configurations/metadata so the scores can be reported properly; iv) evaluation of results by sentences, comparing low or high scores qualitatively. Discussing these characteristics allows students to consider pertinent elements in reproducibility and score reporting and also to investigate how these results can be correlated to HE.

Task 5 - The next task involves one document-level AEM, where students test BlonDe. The exercise with the use of BlonDe is performed through a Jupyter Notebook[1], containing a step-by-step tutorial. In this Jupyter Notebook, students are presented with four steps: i) installing the BlonDe package; ii) uploading text files (the same files from Task 4) in an "input" folder in their academic Google Drive accounts; iii) defining specific code functions to result in an "export" folder that shows the results; iv) running the code to successfully show the results. Using this approach, the teacher should show students how to interpret the results, as the metric measures "Tense", "Pronoun", "Entity", and "Discourse Markers" using Recall, Precision, and F1 scores. Based on the results, the teacher should explain that a higher rating in a specific feature such as Pronoun or Tense may be preferable depending on the context, but to look at the scores considering the task and the purpose of the MT system.

[1]The Jupyter Notebook is available at: <https://colab.research.google.com/drive/1IIJy6FwYnnVzgHE-IZBg7bTRrkd7ycSF?usp=sharing>

Task 6 - Taking a step back from the metrics, the last task of this lesson focuses on measuring the agreement of the HE scores. The teacher should remind the students that in Lesson 2, the first lab session, only Unweighted Cohen's Kappa was tested, and in this class Fleiss' Kappa (Fleiss, 1971), Krippendorff's Alpha (Krippendorff, 1970) and Inter-Rater Agreement are going to be computed. Following the same approach as in Task 5, this is taught using a Jupyter notebook. [2] In this task students follow three steps: i) integrating their Google Drive with the Jupyter Notebook; ii) organising the files by downloading templates and assigning their correct name in the Google Drive structure, as it impacts the code used and; iii) using the templates according to the code made available to them. Students are encouraged to test for each type of measurement. As a guide, to check the calculations, in the Jupyter Notebook a section titled "Practice" has suggestions for students to calculate: i) Weighted Cohen's Kappa with two annotators, on a scale 1-4; ii) Fleiss' Kappa with three annotators, on a Scale 1-5; iii) Krippendorff's Alpha with three annotators, on a scale 1-100.

5.3.3 Assessment of Unit 2

Assessment of this unit is performed after Lesson 4 - the lab session - and is included in the lesson itself as the last task. The assessment should be carried out in the classroom and consists of a short reflective written document of 200 to 500 words. This reflective report focuses on the central theme of adhering to document-level evaluation as an additional dimension to be included in HE approaches and AEMs with the following questions:

1. When performing document-level evaluation, what type of evaluation was more useful to you, document-span or document-level?

2. When performing evaluation, you had different scales (1-4, 1-5, 1-100) and dif-

[2] The Jupyter Notebook is available at: `https://colab.research.google.com/drive/1gWS7wYV3qQRX2X491LbLK9oc4OIcgXX2?usp=sharing`

ferent settings. Amongst these, what scale/setting could be useful to you and what could make things more difficult? (You can think from the perspective of the evaluator, from analysing the data, etc.)

3. When interpreting automatic evaluation metrics, which type was more useful for correlating with human evaluation, sentence-level or document-level?

4. Amongst the automatic evaluation metrics, which seemed to be more useful to you, the sentence-level metrics or the document-level metric BlonDe?

5. How did you feel about the results of different types of measure for Inter-Annotator Agreement?

6. Will you implement any of these approaches/metrics to your practice of MT evaluation?

These questions allow students to compare their experience with the previous lab, making them reflect on the new approaches and metrics they have learned.

Question 1 makes students reflect on the extent to which the situations in which both approaches are useful to them.

Question 2 makes students reflect on two aspects in MTQA: the first being the user's experience when using these various scales and how helpful or unhelpful certain settings can be, and the second focused on their use when controlling biases during the rating of sentences, for example forcing evaluators to make decisions on scales without middle points.

Questions 3 and 4 aims to challenge the students' understanding of the usefulness of AEMs. They have to reflect on the advantages and disadvantages of using newer metrics that may have provided more information than the sentence-level metrics they have used previously.

Question 5 aims to examine whether students' expectations of measuring results with more evaluators in the process would make them question the results of their own evaluations and what could contribute to the change of responses between evaluators.

Question 6 is used to gather additional evidence on whether experimentation with these tools would make students more willing to adopt different approaches in HE and even different types of AEMs, if they were not familiar with document-level AEMs already.

5.4 Unit 3 - Evaluation Design

This unit is designed to consolidate the knowledge of the metrics that students would acquire in Units 1 and 2, and is also a relevant topic suggested by the MTQA lecturers in Phase 1c (described in Section 4.1.2). Toral et al. (2018) explain that testing existing systems that contain open documentation allows verification of claims and validity of results. Läubli et al. (2018) mention that it is important to test new methodologies and evaluation protocols to ensure the validity of an evaluation, as Läubli et al. (2020) argue that we must constantly re-evaluate the design of HE in MTQA. Knowles and Lo (2024) demonstrate through calibration set analyses that different design settings should be tested for evaluations, such as different task instructions or more/less direction for annotators. Because of this, an entire design unit is relevant for NLP students. Thus, this subsection presents the content of Unit 3 (5.4.1), the teaching methodology (Section 5.4.2) and the assessment (Section 5.4.3), following the structure seen in Figure 5.25.

5.4.1 Content of Unit 3

The content of evaluation design follows the research recommendations cited earlier (Toral et al., 2018; Läubli et al., 2018, 2020; Knowles and Lo, 2024), in addition to other guidelines presented in the field, such as Graham et al. (2019). As relevant content, sections from the methodology of WMT finding papers should be presented to the students, such as findings from post-editing (Bhattacharyya et al., 2023) and findings from LLM evaluation (Kocmi et al., 2023). The inclusion of studies using different systems other than NMT is notable, as LLMs become increasingly more

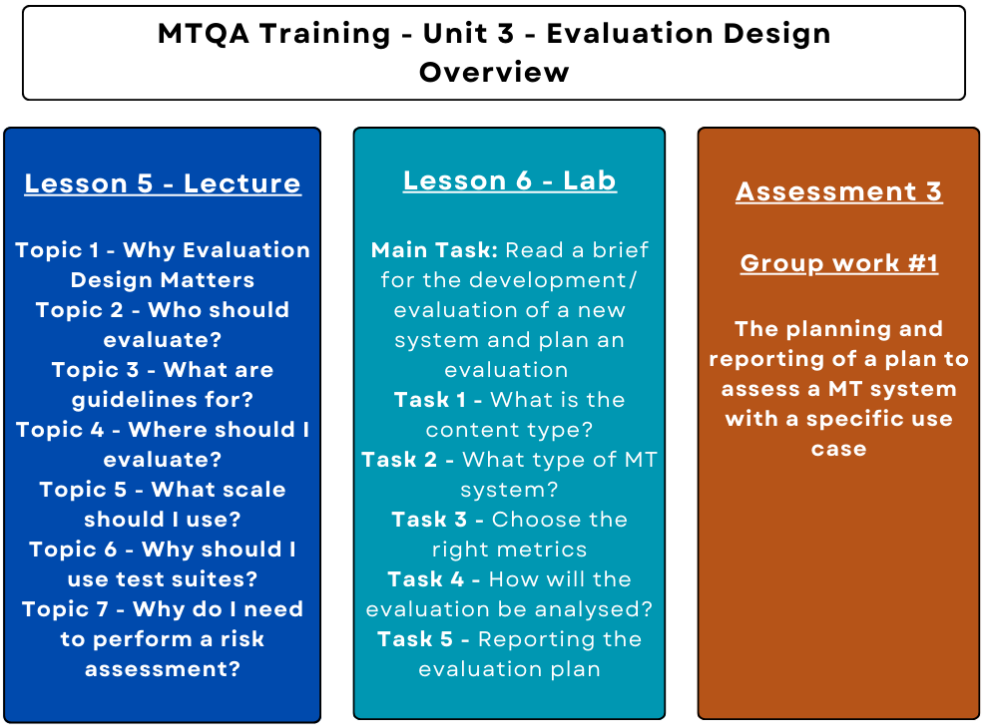


Figure 5.25: Simplified description of topics covered, learning activities and assessment of Unit 3.

relevant and part of shared tasks. Other elements pertinent to the evaluation as factors were considered to be shown to the students, based on the results of the Phase 1c survey, such as the following:

(a) Purpose of MT systems - (Topic 2, Task 1, Task 1a, Task 2, Task 3, Task 3a, Task 3b, Task 4)

(b) Evaluators - (Topic 3, Task 3c)

(c) Guidelines - (Topic 4, Task 3f)

(d) Evaluation Platforms - (Topic 5, Task 3d)

(e) Likert Scales - (Topic 6, Task 3e)

(f) Test suites - (Topic 7, Task 1c)

(g) Risk Assessment - (Topic 8, Task 1a)

These elements are not necessarily new and have been briefly discussed or commented on with students as factors in evaluation. The **(a) Purpose of MT systems** had been discussed previously, but may very well be the starting point of planning any evaluation, whether the purpose of the system is being decided or if it is a routine check on the performance of the system. **(b) Evaluators** are consistently shown or mentioned in Unit 1 and 2, particularly translators as experts who are able to provide an in-depth explanation of linguistic phenomena. International standards are presented to help students use standard criteria to select the right evaluator (ISO, 2015; ASTM International, 2023). (c) **Guidelines** are mentioned, as in previous tasks students would probably ask about instructions or directions, but only in this lesson is the concept of guidelines explicitly discussed in the evaluation process, something that requires attention as discussed in Knowles and Lo (2024). (d) **Evaluation Platforms** are presented to students, given their relevance in UX, which has been a factor discussed in Phase 1c by MTQA teachers and studies such as Briva-Iglesias et al. (2023) which investigated the interaction of translators and MT. Thus, showing platforms and how they can be fit for different evaluators is adequate for NLP students planning evaluations. (e) **Likert Scales** are tied to the previous topic, when HE is used in a given platform. In particular, the way it is measured affects the evaluation, as has been discussed over different tasks in Lesson 2 and Lesson 4. (f) **Test suites** can also be included due to their challenging nature that may align with the objectives of the evaluation, especially since there has been more effort to create test suites for context-level phenomena (Castilho and Knowles, 2024). (g) Risk Assessment is informed both by the international standards mentioned previously and as an aspect investigated in TS, particularly from the studies of Canfora and Ottmann (2018) and Canfora and Ottmann (2020).

5.4.2 Teaching Methodology of Unit 3

To implement the content of Section 5.4.1, two sessions were designed (Appendix F), with Lesson Plan 5 - Lecture and Lesson Plan 6 - Lab. Each lesson plan will

be described explaining the reasoning behind the content and how it interacts with Lessons 1 through 4. The topics covered in Lesson 5 can be seen in Figure 5.26:

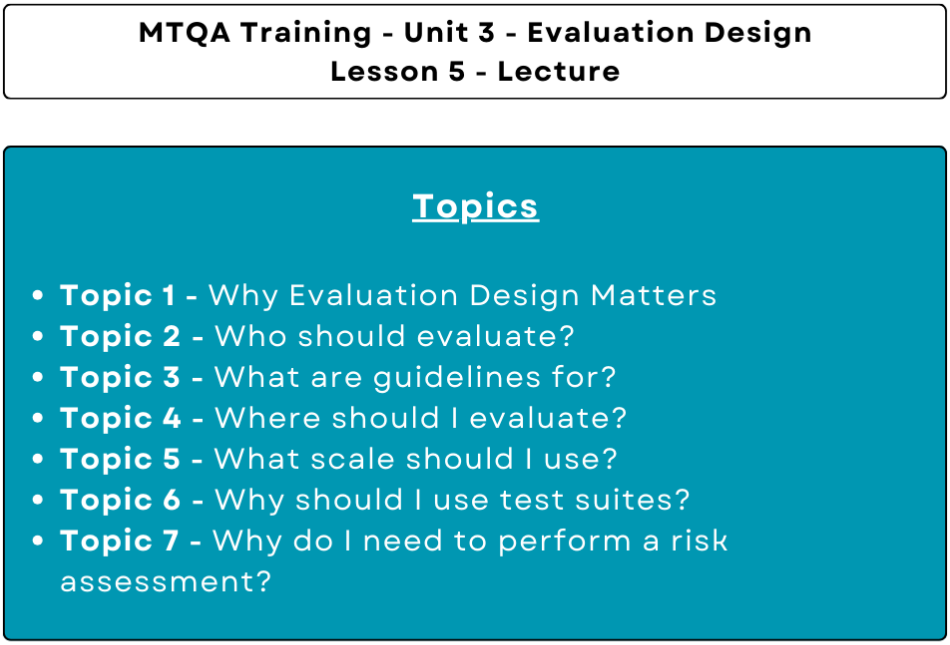


Figure 5.26: Simplified description of topics covered over Lesson 5, delivered as a lecture.

5.4.2.1 Lesson 5 - Lecture

Considering that a reflective report has been assigned in Lesson 4, **Lesson 5** should start with the teacher confirming with students if they have finished the report or not. If not, the teacher has to consider allocating the remaining lesson time for the students to finish that task. The extra time allocated for lecture-based classes would be justified in this approach, for example.

The first topic to be discussed is "Why Evaluation Design matters", making students think why this specific facet of MTQA is relevant. Students should be shown a review of the principles of good MTQA design by Toral et al. (2018); Graham et al. (2019); Läubli et al. (2020). It is an opportunity to extensively explore principles that were practised before, such as whether HE should be done by professional translators and the evaluation methodology should consider the whole

document (Toral et al., 2018; Läubli et al., 2020); claims should be only made for the language pairs and domains tested and should not be generalised (Graham et al., 2019). Excerpts from from WMT 2023 findings papers are used to be examined (Bhattacharyya et al., 2023; Kocmi et al., 2023), specifically from the methodology section. The teacher should project some parts of the methodology from the articles cited for students to identify any good practices implemented in the design of the shared task. It is a good exercise to view how theory applies to practice. As students identify elements, the teacher should move the discussion to outline all elements to be covered in the lesson, such as evaluators, evaluation platforms, guidelines and risk assessment.

The second topic to be addressed is the choice of evaluators. As finding professional translators is essential for evaluation (Läubli et al., 2020), it might be difficult to recruit the ideal evaluators (Doherty, 2017). Using international standards such as ISO (2015) or ASTM International (2023) helps identify criteria widely used and accepted in industry settings. ISO (2015) aims to define professionals as those with a degree or years of experience, while ASTM International (2023) is more broad, allowing SMEs and third-party evaluators for different purposes. SMEs, for example, can be used for monolingual reviews, terminological, or cultural checks.

The third topic revolves around guidelines, as they can be a difficult aspect in MTQA (Doherty, 2017). Guidelines can affect the validity, reliability, and efficiency of the evaluation, according to Knowles and Lo (2024). As the behaviour of the annotator can lead to differences in results, each task in the MTQA must include instructions or guidelines. The teacher should show students that their performance should be monitored for calibration purposes, through levels of fatigue or interaction with the platform (Knowles and Lo, 2024).

The fourth topic should be led by evaluators' UX and the platform where the evaluation happens. Since the way evaluators interact with platforms or sliders in Likert scales affects the evaluation, different platforms must be considered depending on the scenario. The study of Heer and Bostock (2010) demonstrated that anno-

tators had different experiences with platforms depending on the position, length, and colour of the information, including how some elements based on position, such as sliders, can impact annotation processes. Briva-Iglesias et al. (2023) applied the concept of UX to an MT platform for translators using semi-automatic evaluation through post-editing, and a different way of operationalising this process led to higher satisfaction. Because of this, the teacher should present the UI of: Appraise (Federmann, 2018), KantanAI [3], Amazon Turk [4], MATEO (Vanroy et al., 2023) and MutNMT (Ramírez-Sánchez, 2023). The teacher must note that each platform presented has its specific purposes and is not necessarily able to fit all types of MTQA processes - as an example, MutNMT can presented as a friendly tool for translators to learn about MTQA, for example, and this type of tool should be relied on for good UX purposes, not necessarily its functionality to assess the capabilities of an MT system.

The fifth topic focuses on Likert scales and the usability of scales in general. Students experimented with Likert scales in Lesson 2 and with DA scales (used in shared tasks (Graham et al., 2013)) in Lesson 4. Using these examples, the teacher should mention that the UI and type of scale adopted are factors that must also be planned. Not just in terms of UX, but they can reveal patterns. The teacher must explain about the behaviours of evaluators using scales, such as: a) wide distributions (where scoring happens anywhere from 0 to 100), narrow distributions (scoring happens between specific scores, such as 50 and 70), or discretisation (evaluators using always the same scores, such as choosing only 0, 50 and 90 throughout the evaluation). Thus, the teacher advises the students to teach evaluators: how the slider(s) used work; how sensitive the slider(s) must be, whether a mouse or a finger is being used by the evaluators; how many tick marks and labels to adopt in the scale(s); and even if there will be a starting score space, starting at 0, the middle point or any other point to minimise bias. It is important for the teacher to show examples of scales adopted in different tasks from Units 2 and 4 for students

[3]Originally KantanLQR (Luri, 2016), now KantanAI available at https://www.kantanai.io/.
[4]A platform used mostly for Crowdsourcing of services available at https://www.mturk.com/

to analyse.

The sixth topic centres on test suites. As challenging test sets to check how MT systems approach certain linguistic phenomena, evaluation planning can use them as a valuable resource. Although scarce for some phenomena, such as context-level issues (Castilho and Knowles, 2024), MTQA professionals can focus on certain language pairs or issues to test MT systems. To explain further, the teacher should show the DELA corpus, its annotation methodology, and a decision tree that helped the annotation process (Castilho et al., 2021). In addition, the teacher also should share a list of some available test suites (Bawden et al., 2018; Guillou et al., 2018; Müller et al., 2018; Voita et al., 2019).

Lastly, the seventh topic discusses risk assessment and its importance. ISO (2018) is cited as a relevant international standard on the risk management of organisations, and that also applies to MTQA. In addition, the teacher should refer to Canfora and Ottmann's (2020) publication to discuss the issues of NMT from a risk management perspective, showing strategies to avoid, take, share, and reduce risk. The teacher then should explain the type of translation risks that can occur: i) injury or death; ii) legal consequences; iii) loss of reputation; iv) financial damage; v) damage to property; vi) impaired or impossible communication (Canfora and Ottmann, 2018). Because of those risks, the teacher should emphasise the importance of checking for potential risks not just in the output of MT systems, but also test sets that use corpora with errors, as they may propagate errors in the MT system.

The teacher concludes the session by asking students about their thoughts, particularly about any elements of Lessons 1 through 4 they would consider in their planning, and also taking any questions students may have. The teacher should also remind the students to finish the reflective report of Lesson 5. It might be important to note to students that taking longer to complete the report can result in a loss of details. This loss of details might detract from the creation of a insightful reflection report.

5.4.2.2 Lesson 6 - Lab

This lab session uses all the concepts introduced in Lesson 6 to test students' abilities to come up with the design of a MTQA plan. As shown in Figure 5.25, the students have a main task with smaller tasks leading to its completion. Students should be organised in groups and must be given a sheet with a checklist and designed spaces for note taking. Students are supposed to read the evaluation briefs and plan an evaluation of the MT system accordingly. Two briefs are planned. MT System Brief 1 focuses on an MT system capable of handling technical specifications and legal agreements. MT System Brief 2 focuses on an MT system capable of handling user-generated content.

Task 1 - The first task students have to complete is to determine the type of content and its perishability. The decision to include this task follows a recommendation from Doherty et al. (2018) of including this aspect in MTQA education.

Task 1a - Depending on the type of content, students are also tasked with potential risks arising from this type of content, which was discussed in the previous lesson.

Task 1b - Accordingly, the students determine the type of test set used for this type of content, looking at corpus repositories for that choice.

Task 1c - And if relevant, students respond to the question "Will you use a test suite?" to determine if there are any complex linguistic phenomena that they would like to test their system on.

Task 2 - In this task, students choose commercial and freely available MT systems designed for a similar task or purpose, to benchmark their evaluation.

Task 3 - The students must choose the appropriate HE and AEM metrics, aligning the type of metrics with the use case of the MT system, whether it is an established

MT system or if it is a new content area for the system.

Task 3a - HE metrics are chosen at this step, and students have to determine the type of HE metric depending on the use case and how much context would be necessary to evaluate the system.

Task 3b - AEM metrics are chosen at this stage, and students have to choose whether they would use either a combination of reference-based, reference-free, or document-level AEMs.

Task 3c - Accordingly, if using HE, students have to choose the evaluators, employing translators as evaluators and other pertinent stakeholders, whether they are SMEs or end users of the system, or through crowdsourcing.

Task 3d - The evaluation platform is chosen at this stage, and students have to choose from cloud-based survey tools such as Google Forms, Qualtrics, specialised evaluation platforms such as Appraise, KantanAI, or crowdsourcing options such as Amazon Mechanical Turk.

Task 3e - If using HE, students choose how Likert scales should be used and to determine different settings, such as DA scales, their increments, labels, custom positions, values shown to evaluators or snapping the scale at certain values to standardise the scores of the scales.

Task 3f - If using HE, students must decide how the guidelines would be formulated. Students are not required to come up with the complete guidelines as they would have a very short time, but must choose or write down general strategies and approaches to be adopted such as strategies to decrease complexity, offering examples, etc.

Task 4 - To analyse the results, the students have to rely on statistical tests. They decide on statistical tests for IAA measures, statistical tests for correlation of

results, and the analysis of statistical significance, which will be discussed more in Lesson 7.

5.4.3 Assessment of Unit 3

Students are assessed on this task from both their notes and a presentation. Students must report the brief they received and their decisions, task by task, while the teacher takes notes and asks questions. Students have to be qualitatively assessed from both group work and the reporting of their decisions.

5.5 Unit 4 - Ethics in Evaluation

This unit is designed to make students reflect on MT and the implications of (not) performing an adequate evaluation using ethics. Moorkens (2022) warns that evaluations based purely on AEMs that make hyperbolic claims are damaging for translators. Furthermore, because NMT produces errors (Koehn and Knowles, 2017) which bring risks (Canfora and Ottmann, 2018), developers must be trained to bear the responsibility of developing AI systems (Dignum, 2020). People, experts or not, are affected by the development of AI systems as their data are valuable for the development of systems (Moorkens, 2022) and have little choice in whether or not their data are used for that purpose (Moorkens et al., 2024a). In addition to being data hungry, machine learning systems that use the transformer architecture create great environmental impact with carbon emissions (Strubell et al., 2019). Thus, believing that these topics must not be ignored, this subsection presents the content of Unit 4 (5.5.1), the teaching methodology (Section 5.4.2) and the assessment (section 5.4.3), following the structure seen in Figure 5.2.

5.5.1 Content of Unit 4

(a) Addressing the hype (Topic 1, Topic 2, Task 1).

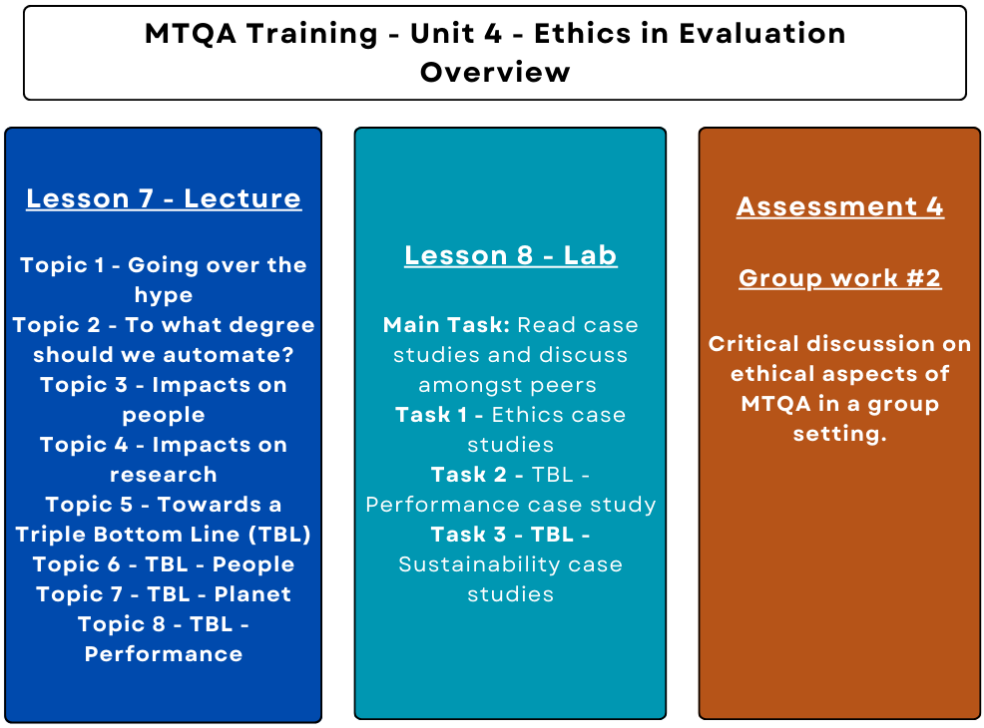


Figure 5.27: Simplified description of topics covered, learning activities and assessment of Unit 4.

(b) To what degree should automation be applied? (Topic 3, Task 1).

(c) Impacts of automation on people: Ethical considerations for translators (Topic 4).

(d) Impacts of automation on research: Ethical considerations for MT research (Topic 5, Task 1).

(e) Towards a Triple Bottom Line (TBL) (Topic 6, Topic 7, Topic 8, Topic 9, Task 2, Task 3).

(a) Addressing the concept of hype starts by addressing the assumption that MT systems can replace translators. The counterargument provided by Moorkens et al. (2024b) is that it is not adequate to claim that MT is a solved problem, as although there has been a considerable amount of work on tools for translation, they are still limited to English and to a lesser extent other languages. Moorkens et al. (2024b) cite Terribile (2024), whose work demonstrates that when edits are

made to MT output, 40% of these edits are considerable. That proportion of edits is considerable, and should not warrant the status of 'solved". There is still a lot to understand about MT systems, thus, reinforcing Moorkens et al.'s (2024b) claim that we, as a community, still need to understand how these models work and eventually be able to explain how they work, as a desirable goal. Part of the reason why these claims exist has to do with automation as a concept, which provides benefits but needs to be rightfully measured and discussed.

(b) To understand to what degree automation should be applied, automation frameworks such as these provided by Parasuraman et al.'s (2000) and Vagia et al.'s (2016) should be shown as examples to students in order to demonstrate the levels of automation and the involvement of the human is when interacting with a system. For contextualisation, the teacher can show an adapted model from TS (Paulsen Christensen et al., 2022), or provide an example using one of the original taxonomies such as that of Vagia et al.'s (2016), as in Figure 5.28.

Level of Autonomy	Description	Type of Translation
Level 1	**Manual control** Computer offers no assistance	a) Human Translation
Level 2	**Decision proposal stage** Computer suggests decisions, operator selects and executes	a) Human Translation aided by TM b) Human Translation aided by CAT tools
Level 3	**Human decision select stage** Human selects a decision, computer executes	a) Human Translation aided by TM b) Human Translation aided by CAT tools c) Post-Editing of Adaptive Machine Translation
Level 4	**Computer decision select stage** Computer selects decision, executes with human approval	a) Human Translation aided by Translation Memories b) Human Translation aided by Computer Assisted Tools c) Post-Editing of Adaptive Machine Translation d) Light Post-Editing of Machine Translation e) Heavy Post-Editing of Machine Translation f) Light/Heavy Post-Editing of MT + CAT tool
Level 5	**Computer execution and human info** Computer executes, informs human	
Level 6	**Computer execution and on-call human info** Executes, informs human only if asked	
Level 7	**Computer execution and voluntary info** Executes, informs only if needed	
Level 8	**Autonomous control** Computer does everything, informs only in case of error	a) Raw Machine Translation

Figure 5.28: Pedagogical model of translation process automation, illustrating eight levels of increasing automation and associated translation types, from fully manual human translation to raw machine translation.

To understand **(c) the impact of automation on people**, the different levels of automation should be linked to a discussion about how it affects not just

the act of translating, but the working conditions and payment. Citing Moorkens and Guerberof-Arenas (2024c) is essential, as they explore an issue that automation brings, namely that it is quite difficult to measure how much MT is being used in a translation. Furthermore, how different types of data are tracked to the translator affects how they are offered jobs, an increasing issue in the age of algorithms, reducing translators' performance and pay to be tied to algorithms that control their agency and worth (Moorkens, 2024).

In view of this, **(d) to understand the impact of automation on MT research**, the next appropriate content to discuss with students should be a direct examination of the ethics applied to MT, through Moorkens (2022), who describes how data is owned, used, distributed. Moorkens (2022) also discusses an issue that is extended to research, which is the overreliance on AEM scores or crowd evaluation to make unwarranted claims about what a system can do.

Adding to that, research in MT presents other problems. As Moorkens et al. (2024b) recommend, AI systems will not only be explainable but must offer layers of explainability to achieve that. For this reason, the research article published by Marie et al. (2021) identifies flaws in MTQA papers published between 2010 and 2020. Discussing this paper with students is relevant, as identified issues can be barriers to rigorous and comprehensive MTQA. These issues include relying on BLEU, not testing statistical significance, copying results instead of reproducing experiments, and making comparisons without factoring in the preprocessing of datasets (Marie et al., 2021).

Lastly, the main goal is to discuss sustainability by proposing the **(e) Triple Bottom Line** (Moorkens et al., 2024a). To do so, the teacher should start talking about the role of translation technology in the human ecosystem, for humans to coexist and co-act with technology (Cronin, 2019). The main purpose is to highlight, from both an environmental perspective, where desktops, laptops, and data centres contribute to carbon emissions, and from an extractivist perspective, where people exploit material resources and the unpaid labour of translators and non-

skilled bilinguals. These factors need to highlight that a healthy ecosystem must be constructed, and this is where the TBL should be mentioned and presented to the students. Based on Elkington's (1997) framework, the TBL is demonstrated to be a repurposed framework to form this healthy co-acting ecosystem that Cronin (2019) suggested, to add translation technology as an essential figure. The idea of the TBL is to find a solution that enables the consistent production of "performance" (adapted from Elkington's (1997) "profit"), while balancing the needs of "people" (their data rights, fair wages, their perspective) and the "planet" (optimising machine learning techniques for more sustainable models).

5.5.2 Teaching Methodology of Unit 4

To implement the content described in Section 5.5.1, two sessions were designed, available in Appendix F, with Lesson Plan 7 - Lecture and Lesson Plan 8 - Lab.

5.5.2.1 Lesson 7 - Lecture

The first topic is "Addressing the hype", where the teacher should encourage students to reflect on the assumption that MT systems can replace translators. The teacher should explain that, although there has been considerable progress in MT, it is not a solved problem, as systems still produce output that requires substantial edits. Students should be led to understand that the NLP community must continue developing a deeper understanding of how MT models function, and that explainability remains a desirable and necessary goal.

The second topic is "To what degree should automation be applied?", where the teacher should introduce students to frameworks that illustrate the levels at which automation can be applied in translation processes. The teacher should use adapted models from the literature to provide context, encouraging students to discuss what the optimal balance between human and machine effort might look like in workflows depending on the use case of a given MT system.

The third topic is "Impacts of automation on people: Ethical considerations for

translators", where the teacher should connect the discussion of automation to its real-world consequences. Students should consider how automation affects translators' working conditions, pay, and agency. The teacher can highlight how data tracking and algorithmic systems influence job allocation and performance evaluation, creating environments where translators' value is increasingly determined by automated systems. This discussion should prompt students to think critically about fairness, control, and ethical labour practices in the translation industry. This should show students how the systems they create impact workers. The fourth topic is "Impacts of automation on research: Ethical considerations for MT research", where the teacher should guide students through the ethical dimensions of MT development and evaluation. The session should cover how data is owned, used, and distributed, as well as the issues arising from overreliance on AEMs and crowdsourced evaluation. The teacher can discuss examples of methodological weaknesses in MT research—such as lack of reproducibility and statistical testing to emphasise the need for transparency, rigour, and ethical responsibility in research design and reporting.

The final topic is the introduction of the TBL, where the teacher should introduce sustainability as a central concept connecting TT to people and the environment. The teacher should explain the three pillars of the TBL framework, performance, people, and planet to encourage students reflect on these aspects. For example, the teacher may ask students if they are aware of the labour practices that led to the development of MT systems they have used in the past. Or, questions about data use and whether they are aware of the data that is collected when they use free MT engines. Finally, the teacher should ask if students are aware of any sustainable approaches or techniques that can be applied to MT development and assessment.

5.5.2.2 Lesson 8 - Lab

Lesson 8 consists of case studies. The first group-based task, involves case studies that contain incorrect, inaccurate, and exaggerated claims about imaginary AI-

driven systems for students to analyse, discuss, correct, and report to the class. The second teacher-centred task, involves the introduction of the TBL framework applied to a study in MTQA, offering students the opportunity to analyse and discuss it from the perspective of "People" and "Planet". The third group-based task involves the analysis and discussion of publications about sustainability in MT and NLP for students to analyse.

Task 1 - For this task, false journalistic reports are created, of systems that supposedly are about to be released. [5]. These reports contain a set of issues that need to be present to test the knowledge students would have acquired from previous lessons and that were highlighted in Lesson 7. To help students, the supposed system name is provided, as was the content type, domain, HE used, and AEMs used. These elements can be omitted if the teacher wishes to provide more challenge. The press release documents contain a short text contextualising the system as if it were being sold, promising absurd capabilities with evaluation practices that are at odds with the literature and would not correlate with the claims being made. In my case, the system names and short texts were initially generated with the help of ChatGPT4o, but I modified them to insert claims that were easily noticeable, evaluation practices that made no sense, and metrics that were incompatible, and added absurd adjectives to make the text hyperbolic. The purpose of this process was to create a text that sounded credible but was fairly inaccurate.

Task 2 - In this task, the teacher provides a regular study reporting the capabilities of MT systems. The students have to identify any aspects that could belong to the elements of the suggested TBL, "People", "Planet", and "Performance". Since I did not want students to potentially criticise the work of peers, I chose to use a paper that I coauthored (Briva-Iglesias et al., 2024). I asked my coauthors if I could adapt our paper for this task, and they accepted it. Due to the short class time for a paper like this, the paper was summarised to contain key elements. The questions

[5]Step-by-step instruction document and activity briefs available at the Github repository

asked are “From this paper, what could fall under the category People?”, “From this paper, what could fall under the category Planet?”, “From this paper, what could fall under the category Performance?” and “What good practices could we apply to this study?”. The key elements were selected carefully so that the students would have enough material to identify and analyse, such as the metrics, approaches, evaluation methodology, language pairs, content type, elements they would be able to identify after Lessons 5 and 6 about evaluation design.

Task 3 - Finally, the teacher provides two papers for students to analyse again in the TBL lens. This time, the papers are explicitly about sustainability, so students would find more TBL elements regarding “People” and “Planet” while explicitly addressing “Performance”. This is designed to make it easier for students to engage as one of the papers is a position paper that advocates for more sustainable initiatives in NLP (Shterionov and Vanmassenhove, 2023) and a paper with a technique that allows reporting of energy and carbon footprints (Henderson et al., 2020). Students would be grouped into two large groups, so they would have the opportunity to debate and discuss.

5.5.3 Assessment of Unit 4

Similarly to the assessment of Lesson 6, the students are assessed by their collaboration. Students are assessed in two ways. For their interactions, discussions, and group work. Each group is asked to document their decisions and observations throughout Tasks 1 and 3. Task 2 is done with the teacher, as it is a preparation for the use of the TBL framework in discussions.

For Tasks 1 and 3, the assessment focuses on the group’s ability to critically engage with the materials provided, identify inconsistencies or aspects of the TBL, and clearly articulate their reasoning. The notes produced by the groups should serve both as a record of their analytical process and as evidence of engagement with the frameworks introduced in earlier lessons, later to be analysed by the teacher.

During group presentations, the teacher takes notes and observes each group's contributions, whether they engage with the core principles of TBL, and how well they collaborate in presenting their views.

Chapter 6

Results & Analysis - Course Delivery - "Implementation"

6.1 Phase 4 - Syllabus Implementation

This chapter is a result of the planned content of Phase 3 (Chapter 5) being implemented into practice with the conditions outlined in Phase 2 (Section 4.3). The purpose of this chapter is to report and interpret where the teaching followed or deviated from the planning. It will emphasise my perspective as a teacher-researcher; therefore, unlike the reporting from the previous chapter, the reporting will return to the adoption of the pronoun "I" and the past tense for the verbs. The conditions in which the course was implemented will be described in Section 6.2. Afterwards, the results will be divided into qualitative (Section 6.3) and quantitative (Section 6.4).

6.2 Contextual Factors in Practice

Following the planning from the design blueprints regarding the participants (Section 4.2.4), I recruited NLP students from the M.Sc in NLP at DCU, a relatively

new programme and the first of its kind in Ireland. [1] They possessed the key characteristic for my doctoral thesis, as **RQ1 - What can be taught to NLP students to encourage them to carry out human evaluation?** specifically required this type of participant. Using the strategy of finding access by communicating with stakeholders (Cohen et al., 2018), I asked one of the lecturers from the master's programme to recruit participants before a lecture in one of the modules they teach. I informed the students of that module about my research and the course details, which were the following:

(a) **Dates:** October 21 - November 11 (4 weeks)

(b) **Schedule:** Tuesdays and Wednesdays, 5:00 PM - 7:00 PM

(c) **Location:** In-person, Dublin City University, Glasnevin Campus, Room GLA:CG08-B

(d) **Total Hours:** 16 hours

(e) **Capacity:** Limited to 20 participants (enrolment on a first-come, first-served basis)

The participants were informed about the content of the syllabus and the LOs, as described in the Chapter 5, Section 5.1. Participants were also informed that the training had a research component and that participation in the research was not mandatory. Participants were reassured that this optional training course would not affect their grade in any mandatory module, as it is not related to their M.Sc. programme. These disclaimers were made to follow the principles of inclusion ethics (Cohen et al., 2018), to decrease the imbalance of power relations the research may have brought into the course. It was also clarified that the benefit of the study would be the course itself and there would be no financial compensation.

Although I set a capacity of 20 participants, I wanted to avoid a low number. The course had discussions and some group work (particularly in Units 3 and 4);

[1] Information retrieved from: https://www.dcu.ie/engineeringandcomputing/news/2023/may/dcu-launches-new-msc-natural-language-processing

therefore, I adopted the homogeneous sampling strategy described in Section 4.2.4. I included PhD students in NLP as participants and students from the M.Sc. in TT. To recruit PhD students, I used a mailing list for members of the ADAPT Centre, as I am affiliated with that research centre. I wrote an email in a tone similar to that of my in-person recruitment, outlining the same details. Moreover, to recruit TT students, I reached out to a lecturer from the M.Sc. in TT also at DCU, asking for a recruitment email to be shared.

Through this recruitment, I reached a total of 19 participants who read and accepted the informed consent form. Out of 19 participants, nine belong to the NLP field, nine belong to the TT field and one participant held a PhD in Education and Literature. In the group of participants from the NLP field, seven were pursuing a M.Sc. in NLP, one was pursuing a PhD in NLP and one was a research assistant in the field of NLP. All TT participants were pursuing a M.Sc in Translation Technology. Their data were pseudo-anonymised for this study and participants will be referred to using their assigned codes starting with P4S (Phase 4 Student) followed by a number, from P4S1 through P4S19.

6.3 Qualitative Results - Practice vs Lesson Plans

The reporting of qualitative results in this section will emphasise the teacher-researcher position from the perspective of reflection. That is, my description and interpretation of events will be done with the aid of research notes I typed throughout and after class. My reporting will focus on deviations from the lesson plan or the success/failure of the practices planned for each topic and task. The lessons presented will include Lessons 1 through 8.

6.3.1 Lesson 1 - Lecture

This lesson followed the lecture component of Unit 1, described in Subsection 5.2.2 (For the topics and learning activities, consult Figure 5.2). Although I set out to have

a dialogic stance, I expected Lesson 1 to be more teacher-centred, as the participants were still getting to know me, and responsiveness and interactions develop over time as described in Section 4.3. I followed Lesson Plan 1 (Appendix F.1) topic by topic, including the distribution of the pretest questionnaire (Appendix G). In the first question I asked the students what they knew about MTQA. I received various responses, ranging from answers such as "Making sure how the model performs", coming from an NLP participant, and "Assessing against a reference translation and seeing how it compares to the output" coming from a TT participant, and "Measure the progress of translation research" as a response from a PhD student. At one point in the lesson, while explaining the use of HE metrics, I was asked about the use of only fluency without adequacy. Following the recommendation of the literature (Castilho et al., 2018; Läubli et al., 2020), I informed students that they typically go together for a more comprehensive evaluation check. Afterwards, the lesson was held according to the lesson plan and no other questions were asked.

6.3.2 Lesson 2 - Lab

Lesson 2 was more student-centred. It had more student comments and interactions throughout the tasks (Figure 5.2 for reference). During **Task 1 - Translators' Perspective, Adequacy/Fluency/Error Classification**, students were constantly making comments. Sometimes they discussed with colleagues or sought confirmation with me. Students asked many questions about specific instructions on what to measure, for example, what would constitute "grammar" to be checked in fluency. I explained to the students that the lack of direction in the evaluation tasks was deliberate and that they should use their own criteria, as guidelines and instructions would be studied in later lessons.

An NLP student asked about the use of human translations in MTQA, I explained that human translations are known as the gold standard. As an example, I explained that they can be included in "ranking" tasks as reference translations, without necessarily making this explicit to the evaluators. I explained that since

the reference translation is the gold standard, the goal is to measure how much the MT output meets the criteria of the task, regarding semantics, grammar, or any other specified criteria that is appropriate for the acceptable standards that that MT system is created for.

Throughout Task 1, a student reported that they did not feel confident doing the task, as they felt it was more appropriate for a professional. Referring to the design of Task 1, another student asked if it is normal to give exclusively low or high scores to segments, when evaluating a system. This question led me to explain that it depends on how the evaluation is set up. The UI contained several segments to be evaluated, but I explained that I could have included only one system, or I could have shuffled the output of several MT systems to decrease bias. I showed my version of the activity sheet on the two screens in the room. It showed that the segments evaluated by the students from Task 1 (see Figure 5.9) were not from one MT system, but from three different ones. I asked students why I might have decided to mix the output of the MT systems across each individual segment. One student answered that it could reduce bias. This reflective moment made the students realise that sometimes the end user can be assessing one system or more than one, and not having that awareness can impact and reduce bias.

In **Task 2 - Inter-Annotator Agreement - Cohen's Kappa**, as students got together in pairs, I walked around the classroom to check on their progress. A pair of students asked what an appropriate level of agreement would be. I reminded them that the agreement varies between each community, but I used the scale adopted by Landis and Koch (1977) to show an example that they could follow.[2] Through this task, the implementation followed according to the development of the course. One NLP student explained to me that because the criteria adopted by them and their peer were different, the results varied wildly. The NLP student suggested that more guidance in the evaluation could lead to a closer level of agreement, which is supported by the literature, as evaluators who are measuring the same things using

[2]See Page 42 of the file "MTQA Course - Slides 1" available at https://github.com/joao-cavalheirocamargo/MTQATraining/blob/main/course_slides.

the same criteria, particularly when it comes to linguistic phenomena (Popović, 2021), tend to have higher agreement when they are trained and instructed on how to evaluate (Lommel et al., 2014). While I noticed that students identified this issue, I asked them how they would approach this problem to achieve a higher agreement. This pair of students responded that they would work extensively on guidelines and perform a pilot evaluation to identify possible issues. I asked other students about their IAA task, if they had reached a low agreement, what possible reasons that could have led to it. I received mixed responses, with some students echoing the same issue of not having guidelines available, while another student pointed out that a low number of sample segments could lead to a more difficult chance of reaching agreement.

In **Task 3 - End user's Perspective: Comprehensibility** and **Task 4 - End user's Perspective: Usability Questionnaire**, students asked why some segments were particularly challenging to evaluate, while others were more or less consistent, as they discussed their decisions and results amongst each other. I explained to students that these instances where it is difficult to properly evaluate may be classified as an edge case, a challenging output that the model produces that leads to extensive discussion among annotators (Klubička and Kelleher, 2024). Another point raised by a student is that it feels less stressful to perform an evaluation as an end user as opposed to from the perspective of a translator. Another student concluded that the type of evaluation also has to be consistent with the evaluation and its purpose, it does not make sense to make a non-expert evaluator post edit if the purpose is to gist, reflecting on the concept of perishability drawn from Doherty et al. (2018). Another student commented that, in their experience, ranking was much easier to perform as there were fewer options to choose.

6.3.3 Lesson 3 - Lecture

This lesson followed the lecture component of Unit 2, described in Subsection 5.3.2 (For the topics and learning activities, consult Figure 5.14). Lesson 3 was more

content-heavy, so students did not engage much directly. While the information was novel, one student commented at one point that it was a lot of information to absorb. This led me to believe that this lesson could perhaps be structured differently. I acknowledge that this potentially goes against a dialogic lecture, as I would identify this lesson as a more traditional, transmissionist lecture. I tried to distance myself from such a stance, and although my intentions were to share a conversation with students, it might be difficult to engage in a conversation when a constant stream of information based on studies can be quite demanding in terms of cognitive load. Even though the stance may be intended to be dialogical, if the design of the lesson is still very transmissionist, it can lead to this incongruence between theory and practice. This led me to believe that it was a weak lesson in pedagogical terms. I hoped that perhaps the active component of the content would provide better results.

6.3.4 Lesson 4 - Lab

In this lesson, my main concern was the number of tasks that the students had to complete. I was aware from the beginning that while the lesson was developed to be comprehensive and cover most aspects of document-level evaluation in HE and AEMs, I also ran the risk of running out of time. I expected the reflective report to be completed only after class, whether on the students' personal time or after Lesson 5.

In **Task 1 - Practice of Metrics - Sentence-level**, already anticipating a difficult lesson, I unexpectedly ran into a technical issue in the first task. When I directed the students to use Qualtrics for the evaluation task, they reported to me that the interface had some sort of issue with settings whereby question numbers were shown twice. It was difficult to follow the instructions, as the instructions would direct the evaluator to a number and the question numbers were randomised. It was excessively difficult to perform the assessment. This issue happened because I imported the questions from one domain into another domain used by DCU as the

university was changing their domain with Qualtrics. My lack of foresight to test the imported questions in the new domain with different settings led to this error. I used my mistake to warn students to always double-test features and interfaces in advance to ensure that no problems arose in the evaluation. If this were to happen in a real evaluation scenario, the results would most likely need to be scrapped, or the quality of the data would decrease significantly. I quickly adjusted the settings and projected the interface to the students so that they could see the questions in the intended way, and requested that they refresh the page to make the changes. That made it possible to complete the task.

In **Task 2 - Practice of Metrics - Context-span** and **Task 3 - Practice of Metrics - Document-level**, students were quite impressed with the interface, as they expressed that it was "friendlier" than using Google Sheets for the evaluation. However, some students noticed that the font sizes and styles were quite different, which led them to express their frustration with it. One student reported that noticing those changes in font led to less satisfaction and made them less enthusiastic about performing the assessment. Students also reported increasing fatigue over the course of the lesson. The exhaustion was understandable for four reasons: i) the time of the class (5 PM to 7 PM); ii) total time of HE tasks (15 minutes each, added up to 45 minutes in total); iii) the analysis of HE approaches and poor UX/UI took time, influenced the tasks and may have increased exhaustion and; iv) evaluating the same sentences across different methodologies could have been boring or exhausting. Because of these four reasons, it led me to think that my approach was flawed, leading students to become more fatigued than expected.

Document-centric approaches are expected to lead to increased fatigue and more time in tasks (Knowles and Lo, 2024), pedagogical factors such as additional tasks could have exacerbated fatigue. Repeated sentences and/or texts in learning activities should be avoided. Although experiencing the negative effects of evaluation is a learning opportunity, I could have warned them formally. Shorter sessions can reduce fatigue and maintain annotation quality (Knowles and Lo, 2024), so shorter

learning activities should be used. Although the UX and UI factors were intentional, the bad experience shows examples of what not to follow. As Briva-Iglesias et al. (2023, p. 8) suggest, it is important to pay more attention to UX in MT and introduce it in various scenarios, including development. For example, in the document-level portions, when DA was set in the platform, I wanted to see the students' reactions and experience with the scales being set up with different intervals, labels, and sensitivity.

In **Task 4 - Sentence-level Automatic Metrics**, I expected NLP students to be familiar with the AEMs [3] and using MATEO (Vanroy et al., 2023) would make this task more linear in calculating the metrics, thus avoiding major technical difficulties. I received mixed results from this task. Both NLP and TT students struggled when navigating the platform, requiring assistance. While it is understandable that TT students would focus more on HE, I had the assumption that NLP students would not struggle with the platform. Although students had a written instruction document available to them, I decided to give some teacher-centred guidance using the projector, showing the platform on my end. The instruction in the lesson plan did not explicitly suggest that the teacher would take a more active approach, but the mixed reactions in the classroom forced me to intervene. This pragmatic approach is in line with the approach of Freire (1996, p. 62-63), where the teacher must respect the students' autonomy, but must also exercise pedagogical judgement intervening when needed. I explained both the process of selecting the right files and the right options in the interface, which could be helpful to students in TT. Afterwards, I explained how to interpret the scores and how to use them in MTQA scenarios.

In **Task 5 - Document-level Automatic Metric - BlonDe**, due to the difficulty, when running sentence-level AEMs, I anticipated similar difficulties for the document-level AEM BlonDe. Although Jupyter Notebooks can be an easier

[3]However, it is possible that both cohorts did not yet formally study the use of AEMs in their modules, as this lesson occurred in the middle of their semester, and this could have impacted their experience with AEMs.

way to introduce computational processes (Krüger, 2022), as students had only recently learnt about AEMs, I was more cautious about this task. Luckily, the only issue in this task involved the management of folders on Google Drive matching the exact name as required in the code, this activity was smoother than using MATEO. Although evaluation platforms and their UI can be user-friendly, a teacher must always be prepared to explain the core principles of the type of evaluation and metric. The Jupyter Notebook containing both step-by-step instructions of how to execute the code and a simple explanation on how to interpret the results allowed students to engage with it more easily. That means that the tool had the right level of scaffolding, as Krüger (2022) suggests to educators designing and developing such a resource for students.

Unfortunately, by this point, the remaining time of the lesson was very short and there was very little time for the Jupyter Notebook focused on IAA measures, meaning that students could not explore it much. While circumstances led me to require more time for AEM tasks due to the sheer amount of content in this lesson, perhaps it would be prudent to diagnose students further in their knowledge of AEMs and distribute AEM tasks in earlier lessons, for example, Lesson 2. In ID, strategies are implemented in the development of a course to reduce an overload of information, and one of them is a better sequencing of lessons, from simple to complex (Seel et al., 2017, p. 79). Upon reflection, that means that this lesson structure has been demonstrated to be difficult for students who are only starting to learn about AEMs, and needs to be reassessed for this type of learner. Although the SRL suggested that more effort must be put towards HE in MTQA teaching, to achieve the recommendation of teaching students to correlate HE results with AEM results, a baseline explanation and understanding of AEMs must also be provided.

I ended the lesson asking students how they felt after the lab session. By their tired demeanour, the students expressed that they associated document-level approaches with fatigue. The students proclaimed that this experience will make them consider fatigue as a factor when planning evaluation for their NLP tasks.

While the literature may support the idea that document-level HE decreases fatigue and increases confidence in evaluation among translators (Castilho, 2021), the application in a pedagogical context can possibly indicate that fatigue is expected when it is a learning task. There is a need for balancing annotator fatigue with document-level features (Castilho and Knowles, 2024), because fatigue can increase time for evaluation and annotations (Knowles and Lo, 2024). The main concern of an inadequate pedagogical approach to document-level methodologies is a decrease in the conditions in which students learn effectively if not scaffolded.

6.3.5 Lesson 5 - Lecture

This lesson followed the lecture component of Unit 3, described in Subsection 5.4.2 (For the topics and learning activities, consult Figure 5.25). I used Vevox to perform a routine check through a quiz (see Appendix F.5.1) on concepts and definitions from Lessons 1, 2, 3 and 4, in addition to a quantitative measurement of students' confidence (described in Subsection 6.4.2) and qualitative measurements. In this lesson, 15 students were present, but not all students responded to all questions, either because they ran out of time or because they had technical issues with the platform. In the quiz, the students were asked to choose the right answer to the question **1 - "What is the purpose of human evaluation?"**, with the following options:

1. To provide subjective opinions without structured guidelines
2. To eliminate the need of automatic evaluation, which is unreliable
3. To provide subjective opinions with structured guidelines
4. To provide insights on complex linguistic phenomena and complement automatic evaluation
5. To provide insights on complex linguistic phenomena and eliminate automatic evaluation, which is unreliable

Twelve students responded to this question. Number (4) was the correct option, and was chosen by 10 students. One student chose number (1) and another student chose number (2). The next question, **2 - "What do reference based automatic evaluation metrics measure?"**, had the following options:

1. The ability to detect tone and emotion in translations
2. The ability to detect tone, emotion, quality in translations between a machine-generated translation and a reference translation (also known as the silver standard)
3. The ability to detect quality in translations between a machine-generated translation
4. The ability to measure similarity between machine-generated translations and reference translations
5. The ability to measure similarity of machine generated translations without a reference translation

Fourteen students responded. Number (4) was the correct option, and was by 13 students. One student chose number (2). The next question, **3 - "What do reference-free automatic evaluation metrics measure?"**, had the following options:

1. The accuracy of a translation in conveying the main ideas, derived from general language models trained without specific reference texts
2. The consistency of style across multiple documents within the same domain
3. The ability of the translation to match the tone and intent of the original text as closely as possible
4. Different aspects of translation quality, depending on the features their model was trained to assess

5. The similarity between the machine-generated translation and a translation generated by a large language model

Twelve students responded. Number (4) was the correct option, and was chosen by three students. Three students chose option (1), two students chose option (2), three students chose option (3), and one student chose option (1). The next question, **4 - "What are the advantages of performing human evaluation at the document level (or with context-span) instead of at the sentence level?"** had the following options:

1. It allows evaluators to consider broader context, but also makes things more ambiguous
2. It helps evaluators ignore minor errors that may disrupt the translation
3. It allows the measurement of adequacy, fluency and comprehensibility
4. It allows evaluators to consider broader context, which can improve assessments by eliminating ambiguity
5. It allows evaluators to consider broader context, which can make assessments quicker

Fifteen students responded. Number (4) was the correct option and was chosen by 13 students. Two students chose option (1). The next question, **5 - "What are the weaknesses of sentence-level evaluation?"**, had the following options:

1. It does not allow you to do group evaluations
2. It makes it easier to overlook spelling and punctuation errors
3. It reduces the evaluator's ability to focus on meaning and focuses too much on style
4. It limits the speed and efficiency of the evaluation process

5. It ignores context across sentences, making the assessment possibly more ambiguous

Fifteen students responded. Number (5) was the correct option. Number (5) was chosen by 15 students. The next question, **6 - "Select the option which only has human evaluation measures"**, had the options:

1. Adequacy, Fluency, BLEU, Comprehensibility, Cohen's Kappa
2. Fluency, Usability, COMET, chrF
3. BLEU, Ranking, Adequacy, Fluency
4. Adequacy, Fluency, Ranking, Usability
5. Adequacy, Fluency, Ranking, Usability, Inter-Annotator Agreement

Fourteen students responded. Number (4) was the correct option, and was chosen by 8 students. Number (5) was chosen by 6 students. The next question, **7 - "What does the concept of perishability refer to?"**, had the options:

1. Texts are meant for immediate consumption, with no purpose afterward
2. Texts meant to be updated regularly
3. Texts that need to be reviewed quickly to maintain accuracy
4. The action of editing machine translated texts
5. The degree to which a text is intended for immediate consumption versus having long-term value.

Fifteen students responded. Number (5) was the correct option, and was chosen by 11 students. Number (1) was chosen by four students. Question **8 - "Rate your confidence level regarding MT quality assessment**" will be described in subsection 6.4.2. Since Lesson 4 was content-heavy on document-level evaluation, I asked students questions to form a word cloud, the first of this type of question



Figure 6.1: Word cloud showing the frequency of terms used in 15 participant responses about the strengths of document-level evaluation, with larger words indicating higher frequency

being **9 - "What are the strengths of document-level evaluation, in your opinion?"**, forming Figure 6.1.

Immediately after, I asked students **10 - "What challenges might arise when performing document-level evaluation, in your opinion?"**, forming Figure 6.2.

Ending the series of questions, I reflected on how well the students had learnt about MTQA so far and how the challenges I faced impacted on their responses and perceptions. The majority of students performed well when checked about concepts related to HE (questions 1, 4, 5), some concepts about AEMs (question 2) and a concept about MTQA design (question 7). However, some questions showed a split in responses, particularly one about HE (question 6) and one about AEMs (3). In question 3, students were unsure about how specifically they could describe a reference-free AEM. I believe that some of the alternatives offered some degree of ambiguity, such as alternative (1) being mostly correct, with the word "accuracy" being used in the sentence as a pitfall to trick students. The point of this question was to measure whether students understood that many reference free metrics are trained on some sort of human judgement, HE metric aspect or a language-specific target feature (Zhao et al., 2020). Question (6), on the other hand, showed that

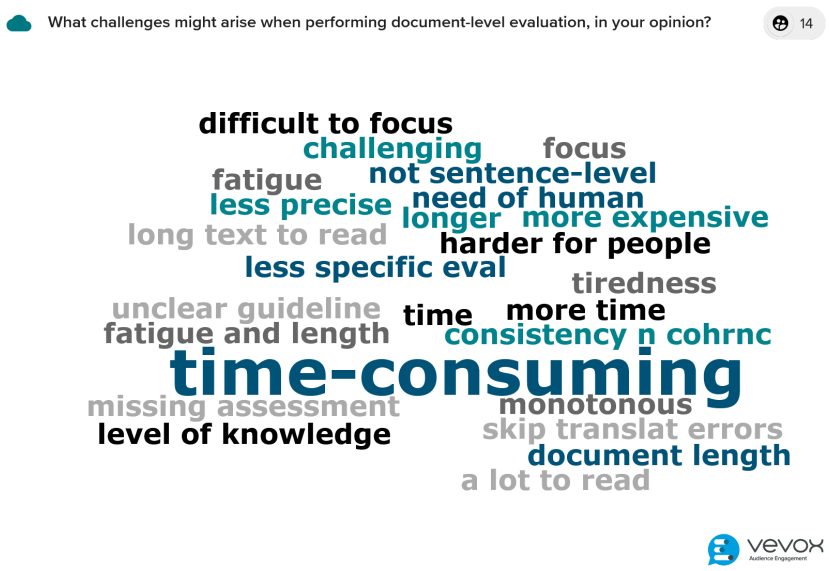


Figure 6.2: Word cloud showing the frequency of terms used in 14 participant responses about the challenges of document-level evaluation, with larger words indicating higher frequency

students still confuse the IAA measures with a HE metric. In Lesson 4, perhaps if more time was allocated to teach AEMs and statistical measures (in the case of IAA), they would be able to identify these pitfalls.

The word clouds from Figures 6.1 and 6.2 allowed me to grasp how students perceive the strengths and weaknesses of document-level methodologies. The word "context" appeared to be a reoccurring theme, and "less ambiguity" is also associated with this approach. Another important aspect identified was "consistent terminology", which is also a feature of document-level evaluation (Castilho and Knowles, 2024). Students' attitudes appear to be positive, as one student felt confident enough to add "cooler" and a thumbs-up emoji. It showed that they felt comfortable enough to demonstrate a positive appreciation for this type of approach with their own words, which is encapsulated in the philosophy of Freire (1996), where the teacher can speak or attempt to understand the student's use of language. If the student is able to convey that information in their own way of speaking, or if the teacher understands how students make meaning with that knowledge in their own language, it shows evidence that lessons were dialogic.

On the other hand, when looking for challenges, "time-consuming", "length",

"monotonous", "a lot to read", "focus" and "fatigue" are mentioned. It showed that some students may have a negative association with document-level evaluation, as the words "monotonous", "difficult to focus", "unclear guideline", "less specific eval" could be related to how they felt when performing the tasks from Lesson 4. The "level of knowledge" mentioned could be associated with the understanding that document-level evaluation is perhaps more beneficial to expert evaluators (Castilho, 2021) and more difficult when no guidelines are provided (Knowles and Lo, 2024). I would not have expected "consistency" and "coherence" to appear as a challenge, as they are, in fact, strengths of document-level evaluation (Castilho and Knowles, 2024, p. 6). Perhaps these could be related to the challenge of having an appropriate setup to achieve that consistency and coherence.

I struggle with the thought that students may have a negative experience with the tasks pertaining to document-level evaluation, but I reflect that they experienced first-hand the difficulty that such a complex process may bring. It makes me think about the ethical dimension of teaching document-level evaluation, as its adverse effects must be acknowledged. Is it acceptable or intended to make students feel exhausted or confused with inappropriate interfaces, or should the focus be shifted to ideal scenarios where nothing goes wrong? These questions, while troubling, still leave room to investigate how the rest of the training unfolded, as this is the halfway point of the course. While the literature suggests that evaluator preferences exist (Castilho and Knowles, 2024; Knowles and Lo, 2024) and UX matters (Briva-Iglesias et al., 2023), a truly dialogic practice of a teacher must instil a sense of firmness (Freire, 1996). Not to be confused with authoritarianism, a proper dialogic practice must be firm, but also it must strive to bring a sense of satisfaction and joy to students in their process of learning (Freire, 1996, p. 139), one that makes them recognise their effort, even if the dialogue is not straightforward or perceived immediately by students. I believe that this human approach is essential to a course that aims to foster the employment of humans in MTQA.

After this Vevox quiz, the lesson on evaluation design proved to be more relevant

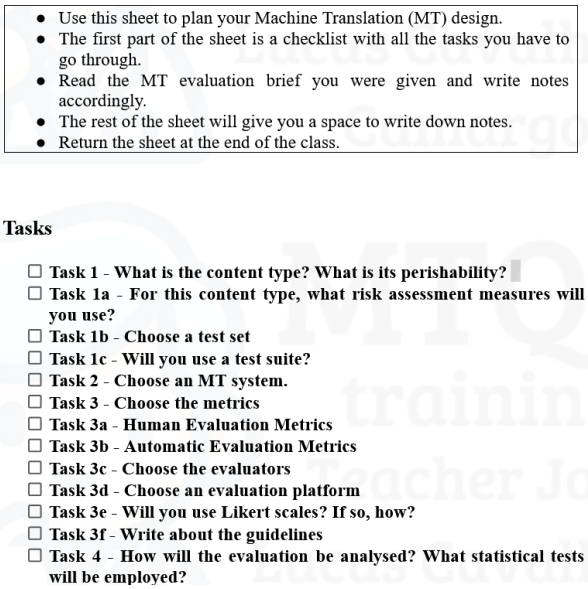
than ever. The gaps in the knowledge of the students could be filled with a lecture by providing a structure: when and where to use each metric, each tool and in what contexts. Most importantly, noticing their shortcomings, I adapted the explanation to address these gaps (Cochran-Smith and Lytle, 2009). Throughout the lesson, I had moments where I asked students to identify good practices by reading excerpts from some papers reporting findings from WMT (Bhattacharyya et al., 2023; Kocmi et al., 2023) and they were successful in identifying the use of expert evaluators, the number of evaluation samples, measures to prevent biases in annotation, evidence of the use of guidelines and the description of the HE metrics used, how the data was collected, assigned, cleaned and corrected. This practical exercise from the lecture made me reflect that their existing knowledge of HE may have contributed to the identification of these aspects from the excerpts.

6.3.6 Lesson 6 - Lab

This lesson put into practice how to structure knowledge built from Lessons 1 to 4 with the resources and frameworks from Lesson 5. The **Main task was to read a brief for the development/evaluation of a new MT system and to plan an evaluation**. The learning activity involved an evaluation plan (Figure 6.3), where I provided a brief (Figure 6.4) to the students, instructions (Figure 6.5), and asked them to fill in the sheets and the checklist as groups.

For this activity to become more interesting from the point of view of group diversity, I assigned the student groups myself, rather than letting them choose their own. Since this lab session would be more group focused, I created groups 1, 2 and 3. Group 1 consisted of an equal split of two NLP students and two TT students. Group 2 consisted only of NLP students. Group 3 consisted of three students in TT and one NLP student. I wanted to see how the different backgrounds interacted in a practical group activity. Students also had access to support materials that comprehensively explained each topic.[4]

[4]Step-by-step instruction document and activity briefs available at the Github repository - https://github.com/joao-cavalheirocamargo/MTQATraining/blob/main/AdditionalMaterials/Lesson%206/.



- Use this sheet to plan your Machine Translation (MT) design.
- The first part of the sheet is a checklist with all the tasks you have to go through.
- Read the MT evaluation brief you were given and write notes accordingly.
- The rest of the sheet will give you a space to write down notes.
- Return the sheet at the end of the class.

Tasks

- ☐ **Task 1 - What is the content type? What is its perishability?**
- ☐ **Task 1a - For this content type, what risk assessment measures will you use?**
- ☐ **Task 1b - Choose a test set**
- ☐ **Task 1c - Will you use a test suite?**
- ☐ **Task 2 - Choose an MT system.**
- ☐ **Task 3 - Choose the metrics**
- ☐ **Task 3a - Human Evaluation Metrics**
- ☐ **Task 3b - Automatic Evaluation Metrics**
- ☐ **Task 3c - Choose the evaluators**
- ☐ **Task 3d - Choose an evaluation platform**
- ☐ **Task 3e - Will you use Likert scales? If so, how?**
- ☐ **Task 3f - Write about the guidelines**
- ☐ **Task 4 - How will the evaluation be analysed? What statistical tests will be employed?**

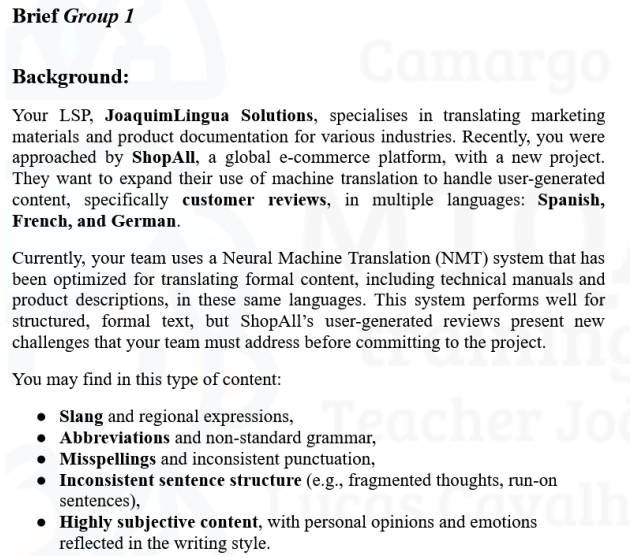
Figure 6.3: Example of the first page of the evaluation sheet, which contained a checklist and note taking spaces for each task

Each brief contained the type of MT system, content, languages or language pairs involved. Although it contained real-life scenarios, it did not have a set timeline or budget to be involved in the brief, as I wanted students to focus fully on how metrics, methodologies, and other decisions interact with each other. Groups 1 and 2 received Brief 1, which dealt with customer reviews and user-generated content. Group 3 received Brief 2, which dealt with legal contracts and technical documentation.

Task 1 - What is the content type? What is its perishability?

- Group 1: "It is user generated content, it is highly perishable".
- Group 2: "Content type: Customer Reviews; Perishability: medium recent time".
- Group 3: "Technical specifications, legal agreements" and "En to Ja, Ko, Ru".

Groups 1 and 2 were successful in identifying both the content type and the perishability, but Group 3 only specified the content type. However, during this process questions arose about the concept of perishability from Group 2, as even though their brief mentioned user-generated content, that content would still be available long term, so that factor was part of their discussion.



Brief *Group 1*

Background:

Your LSP, **JoaquimLingua Solutions**, specialises in translating marketing materials and product documentation for various industries. Recently, you were approached by **ShopAll**, a global e-commerce platform, with a new project. They want to expand their use of machine translation to handle user-generated content, specifically **customer reviews**, in multiple languages: **Spanish, French, and German**.

Currently, your team uses a Neural Machine Translation (NMT) system that has been optimized for translating formal content, including technical manuals and product descriptions, in these same languages. This system performs well for structured, formal text, but ShopAll's user-generated reviews present new challenges that your team must address before committing to the project.

You may find in this type of content:

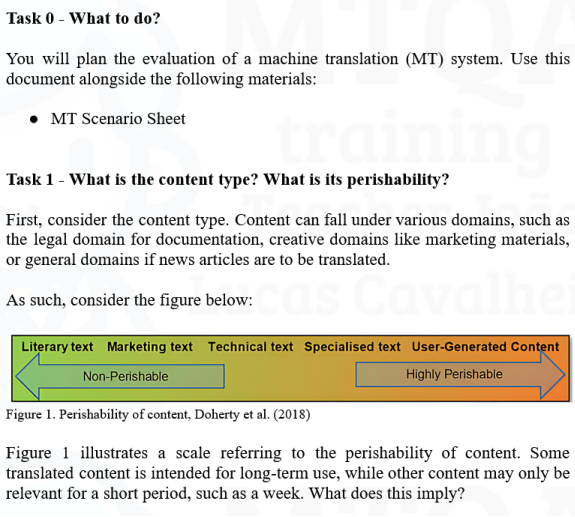
- **Slang** and regional expressions,
- **Abbreviations** and non-standard grammar,
- **Misspellings** and inconsistent punctuation,
- **Inconsistent sentence structure** (e.g., fragmented thoughts, run-on sentences),
- **Highly subjective content**, with personal opinions and emotions reflected in the writing style.

Figure 6.4: Example of part of a brief, which contained a scenario, type of MT system requested for development/evaluation and the context of the current system each group had to adapt

Task 1a - For this content type, what risk assessment measures will you use? :

- Group 1:"Public end-user risk measurement"
- Group 2:"User Experience, Impact of Content, Public or Private use"
- Group 3:"Guidelines, IAA"

Group 1's response seems to consider the ramifications of public end-user data being affected by the type of translation generated by the MT system. Group 2 considers three different factors, the first pertaining to UX, although not many details are given as to how it can reduce risks, the second being related to the impact of the content, which may relate to the type of product and what risks can be associated with a mistranslation, and the last one with the data management, whether public or private. Group 3 only chose to write guidelines and IAA, perhaps with the understanding that a comprehensive and well thought-out evaluation with these elements can mitigate risks.



Task 0 - What to do?

You will plan the evaluation of a machine translation (MT) system. Use this document alongside the following materials:

- MT Scenario Sheet

Task 1 - What is the content type? What is its perishability?

First, consider the content type. Content can fall under various domains, such as the legal domain for documentation, creative domains like marketing materials, or general domains if news articles are to be translated.

As such, consider the figure below:

Figure 1. Perishability of content, Doherty et al. (2018)

Figure 1 illustrates a scale referring to the perishability of content. Some translated content is intended for long-term use, while other content may only be relevant for a short period, such as a week. What does this imply?

Figure 6.5: Example of the first page of the instructions, which provides explanations on what factors to consider for each task in Lesson 6

Task 1b - Choose a test set :

- Group 1:"3 test sets, 1 per language assuming we have 10k customer reviews for each language, 10% will be used as a test set"
- Group 2:"OPUS - Amazon Product Review Dataset"
- Group 3:"1000 sentences in English, references in Japanese, Korean and Russian containing 3000 sentences."

During the course of this task, the students discussed among themselves what would be acceptable. I overheard students from Group 1 discuss how, since their brief mentioned a type of content that has unique features tied to informal language, their test set should consider this. For this task, I advised students not to worry about being too specific, as this was a hypothetical scenario. The purpose of this question was to identify strategies and approaches that students could resort to for their respective scenarios. Group 1, for example, considered the 10% split which is common practice in NLP evaluation. Group 2 focused on the corpus sources, and chose the OPUS dataset (Tiedemann, 2012). I infer that Group 3 presented the final 10% split for the test set.

Task 1c - Will you use a test suite? :

- Group 1:"Yes"
- Group 2:"Yes, build a test suite. Issues: sentiment accuracy, emoji to context, slang / idioms, product info check"
- Group 3:"No"

Group 1 and 3 provided no details in their response. Group 2 brainstormed possible problems that could arise in the deployment of the system, incorporating elements of other NLP tasks. Group 2 considered what Castilho and Knowles (2024) say about how context requires innovation and careful analysis when considering context-level evaluation, which students envisioned with the implementation of other NLP tasks. According to my research diary, I noted that Group 1 planned to use slang as key words to annotate and to train the systems.

Task 2 - Choose an MT system :

- Group 1:"Llama"
- Group 2:"DeepL - For the quality of the system, naturalness and accuracy. Also for the expectation of supporting the language pairs including Spanish, French and German"
- Group 3:"DeepL"

When choosing benchmarks, Group 1 and 2 had a lengthy discussion about the ethical aspects of their brief. Considering that their brief involved a more technical/legal domain, the students were worried whether their benchmarks would be available to use. Because of this, one of the students mentioned that they checked the terms and use of the service for the MT systems they desired to use. In the end, Group 1 chose Llama, which was interesting to witness, as it showed the enthusiasm to use LLMs as a way to explore and compare the quality with that of MT. That

is an example of how LLMs are increasingly used in different ways to examine MT in different contexts (Castilho and Knowles, 2024). Groups 2 and 3 chose DeepL, as it is a well-established MT system. **"Task 3 - Choose the metrics"**, included sub-topics.

Task 3a - Human Evaluation Metrics

- Group 1:"Adequacy/Fluency, Readability (at document level)"
- Group 2:"Usability, Adequacy, Readability"
- Group 3:"Adequacy/Fluency, Usability, Acceptability"

In this task, students were very concerned about what to use, as they wanted to find the appropriate metrics for the type of content and domain. Group 1 mentioned to me that they were having trouble choosing the right methodology due to the length of the content. I showed the pedagogical dataset that I compiled for the Lesson 1 and 2 tasks, which helped contextualise their core issue. Groups 1 and 3 paired adequacy/fluency together, as in a recommendation I made during lessons and suggested in the literature in MTQA (Castilho et al., 2018; Läubli et al., 2018), although Group 2 only mentions adequacy. Surprisingly, Groups 2 and 3 mention usability, as they probably thought about the end-users of the system. It also makes me reflect on whether the usability task from Lesson 2 made the students more willing to adopt this metric in their evaluation plans. Not many details are given on the other metrics, such as acceptability from Group 3 or readability from Group 1 and 2. However, Group 1 did specify the use of a document-level approach, demonstrating that they understand the benefits of HE with more context embedded, and in their discussion, they wanted to factor in time, as they wanted to avoid fatigue in annotation.

Task 3b - Automatic Evaluation Metrics

- Group 1:"Doc-COMET"

- Group 2:"chrF, BLEU, METEOR (versions trained on models) to correlate with humour and new content areas"

- Group 3:"BlonDe"

Group 1 chose Doc-COMET, but offered no rationale for this decision. In addition to being available in the instructions document, it is relevant to recall that this metric was presented in the overview of Lesson 3, so perhaps the choice was inspired by the lesson and the material. Group 2 chose chrF, BLEU, and METEOR, with the rationale that they would use reference free versions trained on models, although I had warned them that these metrics are not trained like COMET or BLEURT. The intention to use AEMs to correlate the results with HE in a particular aspect is an indicator that students have started to think about how to use a combination of HE and AEM, which is desirable. Group 3 chose BlonDe, which was used in Lesson 5. Although not many details were given about their choices, there was one common element. All groups used either a document-level metric or a conversion of a sentence-level metric into a document-level one. This shows that those approaches are seen as valuable and powerful for students to be willing to consider them in their evaluation plans.

Task 3c - Choose the evaluators :

- Group 1:"Translators, active freelancers for Adequacy/Fluency. End-users, for readability"

- Group 2:"Expert Evaluators - To handle complexity of the content. End users - To check user experience (UX)"

- Group 3:"Translators, End-users, Optional - Experts."

In the task related to evaluators, all groups responded with both expert evaluators and end-users. Group 1 specified the type of HE metric according to the evaluator, demonstrating that they understand the distinction of the purpose of

each metric. Group 2 did not specify the type of expert evaluator, but considering adequacy is being used, it can be assumed that the expert is a translator. Group 3 chose translators, end-users and SMEs for specific content, which matches the nature of their brief - unlike Group 1 and 2, Group 3 is working with specialised and technical texts, which brings its own terminology.

Task 3d - Choose the evaluation platform :

- Group 1:"Turk"
- Group 2:"Evaluation - MATEO, Qualtrics, Appraise"
- Group 3:"Google forms"

Group 1 chose to work with crowdfunding for their evaluation. However, it is not specified if this is dedicated for end-user testing, as they mentioned in the previous task for using freelance translators. Group 2 specified different platforms, although not many details are given. MATEO being used as a resource suggests that a tool's familiarity may have made them more willing to use it. As they had two types of evaluators, perhaps it is assumed that for the expert evaluators, Appraise would be used, while for end-user checking, Qualtrics would be adopted. Group 3 mentions only Google forms, but does not provide many details.

Task 3c - Will you use Likert scales? If so, how? :

- Group 1:"We can have a combination, one with yes and no, and Likert Scales for Adequacy/Fluency 1-4"
- Group 2:"Yes, scales 1 to 5. Maintains accuracy, maintains sentiment, ease of analysis"
- Group 3:"1-100 / 1-10 different standard. From worst, bad, middle, good, best."

Group 1 chose "Yes" and "No" options when measuring Readability and Likert Scales for adequacy and fluency, demonstrating how they considered different types of measurement. They provided no additional information, such as what questions or tasks would require the use of a "yes/no" type of response. Furthermore, for adequacy/fluency, a scale of 1-4 was chosen, not having a middle point possibly to force their evaluators to make a choice. Group 2 preferred to use scales 1 to 5, although it is unclear what is meant by their criteria, other than "ease of analysis", which could involve their preference to work with a middle point. Group 3 stated two options, but provided no explanation other than "different standard", and added their own labels.

Task 3f - Write about the guidelines

- Group 1:"We will provide: 1) examples; 2) explain the purpose; 3) tell them what we are expecting from them; 4) give them clear definitions; 5) give them references; 6) explain the platform; 7) explain the structure of the questionnaire; 8) tell the time / parameters"

- Group 2:"Usability, Adequacy, Readability must be clearly defined, ask for clarification"

- Group 3:"We will provide: 1) definitions of the metrics for human evaluation; 2) how to read the Likert scales & answers; 3) outlined the evaluation process; 4) a brief introduction of the project, the purpose, the aim, etc.; 5) document-level evaluation - methodology introduction; 6) privacy declaration"

The question was intended to be generic and broad in order to let the students write what they would prioritise. It is noteworthy that students at some point asked "are we supposed to write the entire guidelines?" and I clarified that they do not have to be so comprehensive, but to focus on a particular aspect they think would be helpful for the guidelines. Groups 1, 2, and 3 focused on providing definitions, explaining what metrics are measuring and how they function. Groups 1 and 2 also

outlined the expectations and criteria in the evaluation, in addition to the purpose of the evaluation and what the project entails. Groups 1 and 3 also emphasised explanations regarding scales and methodology, particularly with document-level evaluation mentioned by Group 3. Group 1 mentioned that the platform would be explained. Other factors include time, mentioned by Group 1, and encouraging evaluators to ask for clarification, mentioned by Group 2. Finally, Group 3 included privacy as an element, which leads me to believe that the students in this group were concerned about the ethical components of an evaluation.

Task 4 - How will the evaluation be analysed? What statistical tests will be employed? :

- Group 1:"We will use the IAA measure, Krippendorf's alpha"
- Group 2:"We have to analyse language pairs, finding from human evaluations. Statistical tests - mean, medium, mode, standard deviation. Cohen's Kappa (weighted), Correlation matrix (Pearson coefficient)"
- Group 3:"IAA - Krippendorf's Alpha"

Considering that the students did not have any datasets at their disposal to analyse or calculate, I understand that this task might have been difficult. Although in the instructions the students had flowcharts to analyse the results, no student mentioned any statistical tests. However, I believe that students have made attempts to imagine scenarios and how they would relate to their own evaluation project.

Each group presented their brief and evaluation plan. I did not provide a structure, but an overall time frame of 10 to 15 minutes, so they could do each task and explain what motivated their decision. The groups divided the presentation amongst the group members, a decision they made among themselves. It was positive to notice their autonomy in making these decisions. After the presentations, I asked about the difficulty of the task, some students reporting that they found it complex and difficult, as it is not something they have thought about before. However,

I consistently received feedback from NLP students that the knowledge from this training has been helping them with their coursework in their M.Sc. programme, as they feel more confident doing activities related to HE and AEMs.

6.3.7 Lesson 7 - Lecture

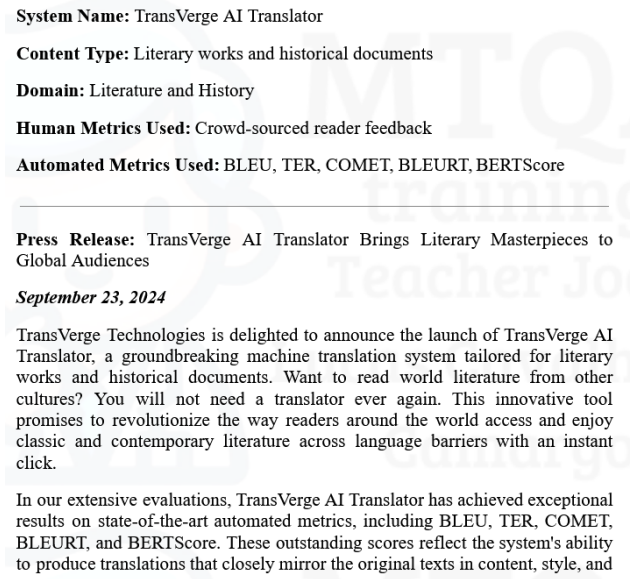
This lesson followed the lecture component of Unit 4, described in Subsection 5.5.2 (For the topics and learning activities, consult Figure 5.27). When I delivered this lesson, I expected it to be a 'heavier' lesson. When the lecture ended, I asked the students if they had any thoughts. One of the students mentioned "This topic, Ethics, is quite complex and I am still thinking about it". Another mentioned "It is really interesting, but it is a lot to process". It gave me relief to receive this feedback, as I thought that the delivery of the content not being engaging was entirely my fault. However, that does not deter me from thinking that I should think about other ways to teach ethics. Although the lesson attempted to be dialogic, I believe that I did not achieve that, as this lesson focused too much on the content and did not provide meaningful options for the students to engage with. However, the labs pertaining to this topic could show students' engagement more directly.

6.3.8 Lesson 8 - Lab

In this lesson, I had three tasks involving case studies. **Task 1 - Ethics Case Studies** were case studies for students to demystify hype, identify flaws in evaluations, and think about ways to promote good, responsible and hype-free reporting. **Task 2 - TBL - Performance Case Study** involved a case study of a paper that falls under "Performance" to verify how students would consider additions of "People" and "Planet" following the TBL. Finally, **Task 3 - TBL - Sustainability Case studies** directly engaged with all aspects of the TBL had students brainstorm ideas to categorise Performance, People and Planet aspects in MTQA.

In **Task 1 - Ethics Case Studies**, I handed the students the ethical case

studies. [5] An example of a case study can be seen in Figure 6.6



System Name: TransVerge AI Translator

Content Type: Literary works and historical documents

Domain: Literature and History

Human Metrics Used: Crowd-sourced reader feedback

Automated Metrics Used: BLEU, TER, COMET, BLEURT, BERTScore

Press Release: TransVerge AI Translator Brings Literary Masterpieces to Global Audiences

September 23, 2024

TransVerge Technologies is delighted to announce the launch of TransVerge AI Translator, a groundbreaking machine translation system tailored for literary works and historical documents. Want to read world literature from other cultures? You will not need a translator ever again. This innovative tool promises to revolutionize the way readers around the world access and enjoy classic and contemporary literature across language barriers with an instant click.

In our extensive evaluations, TransVerge AI Translator has achieved exceptional results on state-of-the-art automated metrics, including BLEU, TER, COMET, BLEURT, and BERTScore. These outstanding scores reflect the system's ability to produce translations that closely mirror the original texts in content, style, and

Figure 6.6: Example of one of the case studies crafted for Task 1, containing preposterous claims, inaccurate reporting or questionable evaluation practices

As I organised the groups and presented the case studies, I could hear different comments from the students, ranging from their initial impressions and when they started to analyse. "This sounds like a load of shite", a student said out loud while reading. Another student looked at me and said "The whole thing is off. This is dodgy, man!", showing that they had immediate suspicion about the claims of the document. Another student related the exercise to their world knowledge: "They are like the politicians of Fianna Fáil, they promise everything", which made me think how my training may have reached students on a Freirean level. The fake briefs that presented the miraculous MT/LLM systems made students think about other aspects of development and assessment. One student told me "I have realised that the more I study about this, the more I study about LLMs, the less I will use it. I do not want to give them my data". Perhaps their in-depth study of MTQA has made them realise the exchanges made as an end-user with those systems when data and user input are precious currencies. It echoes the Freirean thought that the sense of ethics and politics is greater than the technological dimension (Freire,

[5] Step-by-step instruction document and activity briefs available at the Github - https://github.com/joao-cavalheirocamargo/MTQATraining/tree/main/AdditionalMaterials/Lesson%208/

1996, p. 127). As I taught students, I noticed that they were able to "compare, observe, assess" their reality while being ethical individuals (Freire, 1996, p. 98). That is, individuals who recognise that there is no neutrality, and as dimensions in life are political and ethical, they are able to relate that critical thought towards their own expertise. Their willingness to share their thoughts while using their own vocabulary with slang or correlations to their perspective on life has made me realise that I achieved a dialogic lesson when students felt comfortable to be in a position to share their own thoughts.

As the activity progressed, I noticed the students telling me how they are applying the knowledge from the course to their core modules in their M.Sc. One student told me "I think this is wrong because we need to do what one of our professors said. You should release your datasets, your models..." when pointing out sustainability and transparency concerns in development. Another student reflected on their own progress with the course and the activity: "These look so fine and believable. If we had not covered what we covered in this training, we would blindly believe it". That shows evidence of a change of attitude towards dealing with MTQA processes, not just applied to MT, but to other tasks in NLP in general. As the students identified flaws, the students in Case Study 1 were concerned about data collection for the training of the systems. Students in Case Study 2 acknowledge that it is a press release and understand that they would not get details about the claims, but would verify the claims with the methods and results of their evaluation. Students in Case Study 4 were concerned about the lack of detail about specific languages and wondered how the systems they were evaluating would work with low-resource languages.

The following activity was teacher-centred. I read a summary of a paper with the students and asked them to identify which aspects of the TBL would pertain to Performance, People, and the Planet. If no aspects could be identified, how could the study be improved to include those aspects and discussed. However, the students were quiet throughout the activity. I did not receive much feedback. As

the silence continued, it made me question the effectiveness of this task. Making it teacher-centred might have been a mistake. Although the intention was to vary the activities throughout the lesson so that the students did not get tired or bored, the specific design or way of leading this activity could have made the students not fully engage in it. Therefore, I consider it to be a flaw in the development of this task. Students could have received a summarised version of the paper themselves and could have identified these aspects on their own.

The last activity consisted of organising the students in two different groups. In one group, the students read a summarised version of Henderson et al. (2020) and Shterionov and Vanmassenhove (2023), since there was only a short amount of time; to allow students to read the full version, the full reference and link to the papers were provided. The purpose of this learning activity was to introduce papers focusing on sustainability in MTQA and NLP in general to students, in addition to providing the opportunity for students to identify which aspects of those papers would fall under each paradigm of the TBL framework. Group 1 analysed Henderson et al. (2020), and identified the following aspects for People:

TBL Category	**Group 1 – Henderson et al. (2020)**	**Group 2 – Vanmassenhove (2021)**
TBL – People	All AI researchers; the authors; general public	Language pairs; general population; Ireland, Netherlands
TBL – Planet	Carbon emissions; energy consumption; carbon footprint; resource-intensive; kilowatt-hours	Carbon emissions; energy consumption or usage; energy profits; eco-friendly alternatives; environmental cost; Green AI
TBL – Performance	Green metrics → performance + energy consumption; → kilowatt-hours energy consumption; energy-efficient models	Optimisation; trade-off of performance and energy; translation quality; model reusability
What to bring to MTQA?	Transparency in reporting energy use; accessibility	No response

Table 6.1: Student mapping of Triple Bottom Line (TBL) elements in two sustainability papers, with notes on their relevance to MTQA.

When thinking about People, Groups 1 and 2 identified the population as a group that would benefit from the direct output of implementing green strategies

and tools into systems. Group 1 focused on researchers, which is adequate, as the purpose of a position paper is to raise awareness of an issue or trend. Group 2 pointed out the language pairs cited in the paper. While language pair claims must not be generalised to other language pairs (an MTQA recommendation probably informed by Graham et al. (2019)), students perhaps reflected on this when reading the summarisation. However, the technique implemented by Shterionov and Vanmassenhove (2023) was focused on the architectures themselves, so the language pair might not be an impacting factor in this case. Nevertheless, it shows an effort by the group to consider what elements matter when applying the TBL. Another element identified by Group 2 was the specific geographic marker, which shows a perception of scale and granularity of impact across studies, which is an interesting point to note.

Moving to Planet, both groups identified important elements such as carbon emissions and energy consumption. While Group 1 focused on the quantitative framing of ecological impacts, Group 2 emphasised themes that could directly contribute to environmental change, such as eco-friendly alternatives and Green AI. The identification of "energy profits" was also notable, especially when "profit" can be traced back to the original TBL framework (Elkington, 1997).

Finally, when discussing Performance, Group 1 looked at the discussion through green metrics, showing kilowatt-hours as a quantitative element that can be taken into account. Group 2 identified the advantages of the technique introduced by Shterionov and Vanmassenhove (2023), such as optimisation and the trade-off between performance and energy, which impacts translation quality.

When reflecting on the implications of these papers for MTQA, students were asked to brainstorm any approach, implementation, or applicable idea. Although challenging, Group 1 identified transparency in energy reporting and accessibility as transferable values. Group 2 did not explicitly respond, but reported that it was difficult and complex to find a clear solution, as the paper led them to think.

6.4 Quantitative Results - Measuring Effectiveness

To measure effectiveness and complement the qualitative results from the previous section, this section presents the quantitative results. Drawing on the one-group time series design described in Section 4.4.2. Subsections 6.4.1, 6.4.2, 6.4.3 report descriptive statistics at three measurement points: at the beginning of training, the mid point, and after training. Finally, subsection 6.4.4 details the inferential analysis used to assess change and its significance observed at the different points of the course.

6.4.1 Pretest Questionnaire Results

At the pretest stage, when the course started, participants responded to the questionnaire in Appendix G, with responses ranging from 1 to 5 on a scale. Results of questions "9 - How confident are you in using automatic evaluation metrics?" and "10 - How confident are you in using human evaluation metrics?" are in table 6.2, where "HEConfidence" is confidence in using HE and "AEMConfidence" is confidence in using AEM.

Variable	N	Min	Max	Mean	SD	Mode
HEConfidence	19	1.00	5.00	2.11	1.10	1.00
AEMConfidence	19	1.00	5.00	2.37	1.16	3.00

Table 6.2: Descriptive statistics for HE and AEM confidence at Pre-test

Participants reported low confidence levels in both types of evaluation, with the mean score for HEConfidence 2.11 (SD = 1.10) and for AEMConfidence slightly higher (M = 2.37, SD = 1.16). These initial results suggest that while some participants had moderate familiarity with AEMs, confidence in HE evaluation approaches was notably limited across the cohort. To explore these patterns in greater detail, frequency distributions are presented below separately, starting with HE in table 6.3.

Score	Frequency	Percent	Cumulative Percent
1	7	36.8%	36.8%
2	5	26.3%	63.2%
3	6	31.6%	94.7%
4	0	0.0%	94.7%
5	1	5.3%	100.0%
Total	19	100.0%	–

Table 6.3: Frequency distribution of HE confidence scores at Pre-test

The frequency distribution of HE confidence scores concentrate low confidence level among participants. Over a third of the cohort (36.8%) selected the lowest possible rating (1), and nearly two-thirds (63.2%) reported confidence scores of 2 or lower. Only one participant reported a score of 5, and none selected 4, indicating a strong skew towards low perceived confidence in HE metrics. This distribution reinforces the notion that the participants were unfamiliar with or uncertain about HE metrics.

Score	Frequency	Percent	Cumulative Percent
1	6	31.6%	31.6%
2	3	15.8%	47.4%
3	8	42.1%	89.5%
4	1	5.3%	94.7%
5	1	5.3%	100.0%
Total	19	100.0%	–

Table 6.4: Frequency distribution of AEM confidence scores at Pre-test

According to Table 6.4, AEMConfidence scores are slightly more positive compared to HEConfidence. A sizeable proportion of participants (31.6%) still reported the lowest score of 1, a slightly larger share (42.1%) selected a mid-point score of 3, indicating moderate familiarity with AEMs. Only two participants (10.6%) reported high confidence levels (Scores of 4 or 5) suggesting that while the group was somewhat more comfortable with AEMs, high confidence remained uncommon.

6.4.2 Mid-test - Classroom quiz Results

As reported previously in Lesson 5 (subsection 6.3.5), in the Vevox quiz, participants had their confidence measured in one question, also using a scale. It is worth noting that at this point some participants had dropped out or were unavailable, resulting in a reduced sample size. From this point on, the analyses reflect a smaller number of responses, as shown in Table 6.5.

Variable	N	Min	Max	Mean	SD	Mode
HEConfidence	15	2.00	4.00	3.67	0.62	4.00
AEMConfidence	15	2.00	4.00	3.40	0.74	4.00

Table 6.5: Descriptive statistics for HE and AEM confidence at Mid-test

At the mid point, both confidence variables showed notable increases relative to the baseline scores. HEConfidence rose to a mean of 3.67 (SD = 0.62), with the mode shifting from 1 to 4, indicating a marked rise in confidence. Similarly, AEMConfidence increased to a mean of 3.40 (SD = 0.74), also with a modal score of 4. These shifts suggest that participants were beginning to feel more confident in their use of HE metrics and AEM metrics by Lesson 5, reflecting the impact of the initial half of the training programme.

Score	Frequency	Percent	Cumulative Percent
2	1	5.3%	6.7%
3	3	15.8%	26.7%
4	11	57.9%	100.0%
Total (valid)	15	78.9%	–

Table 6.6: Frequency distribution of HE confidence scores at Mid-test

Observing Table 6.6, the frequency distribution of HEConfidence illustrates a pronounced upward shift in self-reported confidence in HE metrics. A majority of the participants (57.9%) selected a score of 4, compared to none at the pretest. Only one participant remained at the lower end of the scale (score of 2), and no responses were below 2. This distribution suggests that, by the midpoint of the training, confidence in using HE metrics had improved substantially for most of the

participants, with group clustering at the upper end of the scale.

Score	Frequency	Percent	Cumulative Percent
2	2	10.5%	13.3%
3	5	26.3%	46.7%
4	8	42.1%	100.0%
Total (valid)	15	78.9%	–

Table 6.7: Frequency distribution of AEM confidence scores at Mid-test

In a similar fashion, as seen in Table 6.7, confidence in using AEMs rose, but showed more variance. The most frequent score was 4 (42.1%), and over two-thirds of participants (68.4%) selected scores of 3 or above. Nevertheless, two participants (10.5%) still reported relatively low confidence at level 2. That could potentially be explained by the issues in Lesson 5 regarding the distribution of time for learning activities involving AEMs, as reported in 6.3.5. The participants who self-assessed a low score could have associated their confidence with the experience they had in the lesson.

6.4.3 Posttest Questionnaire Results

At the posttest stage, when the course ended, participants responded to the questionnaire in Appendix H, in which questions "5 - How confident are you in using automatic evaluation metrics?" and "6 - How confident are you in using human evaluation metrics?" were asked, with responses ranging from 1 to 5 on a scale. These questions elicited the results in Table 6.8.

Variable	N	Min	Max	Mean	SD	Mode
HEConfidence	14	2.00	5.00	3.57	0.76	4.00
AEMConfidence	14	2.00	4.00	3.50	0.65	4.00

Table 6.8: Descriptive statistics for HE and AEM confidence at Post-test

By the end of the course, participants' confidence remained elevated relative to the beginning of the training. The mean HEConfidence score was 3.57 (SD = 0.76), slightly lower than the mid-test mean but still substantially higher than the pretest

level. The mode remained at 4, suggesting continued high self-assessed competence in HE. AEMConfidence followed a similar pattern, with a mean of 3.50 (SD = 0.62) and a mode of 4. These values indicate sustained gains in both domains, with only minor variation in the central tendency compared to the midpoint.

Score	Frequency	Percent	Cumulative Percent
2	1	5.3%	7.1%
3	5	26.3%	42.9%
4	7	36.8%	92.9%
5	1	5.3%	100.0%
Total (valid)	14	73.7%	–

Table 6.9: Frequency distribution of HE confidence scores at Post-test

The frequency distribution for HEConfidence in the posttest in table 6.9 shows a generally sustained level of confidence, although with a slightly broader spread than at the midpoint. The majority of the responses (36.8%) clustered at level 4, and one participant reported the maximum score of 5, indicating strong confidence. However, five participants (26.3%) selected a score of 3, and one reported a score of 2, suggesting a modest regression or recalibration in confidence among some participants. Overall, the distribution suggests that the initial gains in HE confidence were largely maintained, albeit with slight levelling off or individual variation by the course's conclusion.

Score	Frequency	Percent	Cumulative Percent
2	1	5.3%	7.1%
3	5	26.3%	42.9%
4	8	42.1%	100.0%
Total (valid)	14	73.7%	–

Table 6.10: Frequency distribution of AEM confidence scores at Post-test

The AEMConfidence scores in Table 6.10 reflect a continued upward trend, with most participants reporting higher levels of confidence. The majority (42.1%) selected a score of 4, consistent with the mode observed at the midpoint. A further 26.3% chose a score of 3, while only one participant (5.3%) reported a low confi-

dence score of 2. Although no participants selected the maximum score of 5, the overall distribution indicates sustained improvement in confidence relative to the baseline scores. Similar to the confidence in HE, these results suggest that participants generally retained the confidence gains acquired during the earlier phases of the training.

Taking these scores together, the descriptive statistics and frequency distributions across the three timepoints suggest a clear trajectory of increasing confidence in both HE and AEM domains. Participants began the training with relatively low confidence levels, particularly in HE. By the midpoint, there was a marked increase in both mean scores and modal responses, indicating that the early stages of the course had a substantial positive impact. At post-test, these gains were largely maintained, though with some individual variability and slight levelling in HEConfidence. These descriptive trends indicate preliminary evidence of training effectiveness, which will be further examined through inferential analysis in the following section.

6.4.4 Inferential Analysis: Bayesian Linear Mixed Modelling

To examine whether the observed changes in confidence over time were statistically significant and to account for individual-level variation, I chose to use a Bayesian LLM. This subsection focuses on the HEConfidence variable, while full results can be consulted in Appendix H. Fixed effects included Time (Pretest, Mid-test, Posttest), Background (NLP vs TS), and their interaction.

Effect	**Num df**	**Den df**	**F**	**Sig. (p)**
Intercept	1	14.051	516.74	$< .001$
Time	2	25.665	17.73	$< .001$
Background	1	14.051	1.75	.207
Time × Background	2	25.665	0.36	.701

Table 6.11: Summary of fixed effects from the linear mixed model for HEConfidence. The table reports each predictor, the numerator and denominator degrees of freedom used in the F-test, the F-value, and the associated significance level (p-value).

The linear mixed model analysis revealed a significant main effect of Time on

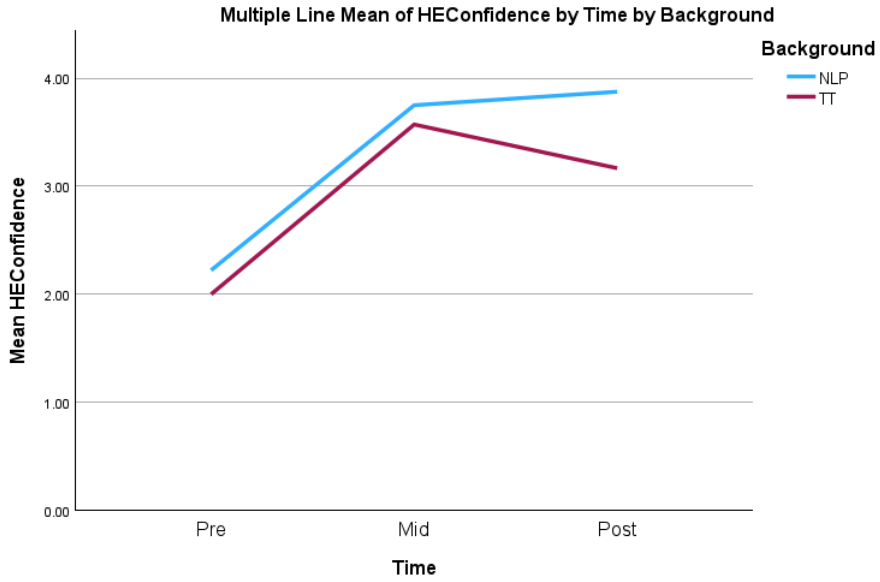


Figure 6.7: Mean *HEConfidence* across time, by participant background (NLP vs TT).

HEConfidence, $F(2, 25.67) = 17.73$, $p < .001$., indicating substantial changes in participants' confidence levels across the three measurement points. In contrast, the main effect of Background ($F(1, 14.05) = 1.75$, $p = .207$) and the Time × Background interaction ($F(2, 25.67) = 0.36$, $p = .701$) were not statistically significant, suggesting that confidence increased similarly regardless of participants' disciplinary background.

Estimates of marginal means (Table 6.12) showed a significant increase in HEConfidence from Pre (M = 2.11, SE = 0.202) to Mid (M = 3.65, SE = 0.227), with a slight decline at Post (M = 3.53, SE = 0.237), although still markedly higher than at baseline. Pairwise contrasts revealed that the Pre–Mid increase was statistically significant (p=.011), while the Mid–Post change was not.

Inspecting Figure 6.7, when broken down by background, participants from both NLP and TT groups followed a similar trend, showing large improvements from Pre to Mid and then levelling off or slightly declining by Post. The NLP group showed marginally higher overall scores, but the non-significant interaction confirms that the pattern of change over time was consistent between the two groups.

In the next stage, calculating the estimated marginal means provides model-

adjusted average confidence scores at all time points. These means offer a clearer picture of how HEConfidence changed throughout the study, accounting for individual variation and the fixed effects included in the model. Table 6.12 presents the estimated means for each timepoint.

Time	Mean	SE	95% CI Lower	Upper
Pre	2.111	0.202	1.704	2.519
Mid	3.653	0.227	3.194	4.112
Post	3.526	0.237	3.048	4.004

Table 6.12: Estimated marginal means of HEConfidence by Time

The estimated marginal means in Table 6.12 show a marked increase in HEConfidence from the Pre-test (M = 2.11, SE = 0.20) to the Mid-test (M = 3.65, SE = 0.23), indicating a substantial gain in confidence during the first half of the training. At Post-test, the mean remained high (M = 3.53, SE = 0.24), suggesting that this increase was maintained through to the end of the course. The relatively narrow confidence intervals across all timepoints further support the stability of these estimates. While there was a slight decrease from Mid to Post, the difference was minimal and within the margin of error.

Chapter 7

Discussion - Integrating Classroom Data into Praxis - "Evaluation"

This chapter, referring to Phase 5, aims to answer RQ1 (What can be taught to NLP students to encourage them to carry out HE?), RQ3 (What can be drawn from TS when teaching MTQA for NLP postgraduate audiences?), and RQ4 (How effective is the content proposed for HE in an MT evaluation module?) by converging the data collected throughout the thesis, including results from Phase 1a (SRL), Phase 1b (Curricular analysis), Phase 1c (Survey), Phase 4 (Implementation). This is performed through the thematic analysis of students' attitudes. To answer RQ1, I focus on students' attitudes and whether their willingness to use and adopt HE approaches increases over the course of the training. To answer RQ3, I focus on students' mentions of resources, methodologies, and metrics stemming from TS. To answer RQ4, I focus on the pedagogical aspects and their effectiveness, and therefore, I check students' abilities to demonstrate that they can accomplish tasks described in the LOs. Data from Phase 4 is triangulated where relevant with data from previous phases (1a, 1b, 1c, 2, 3).

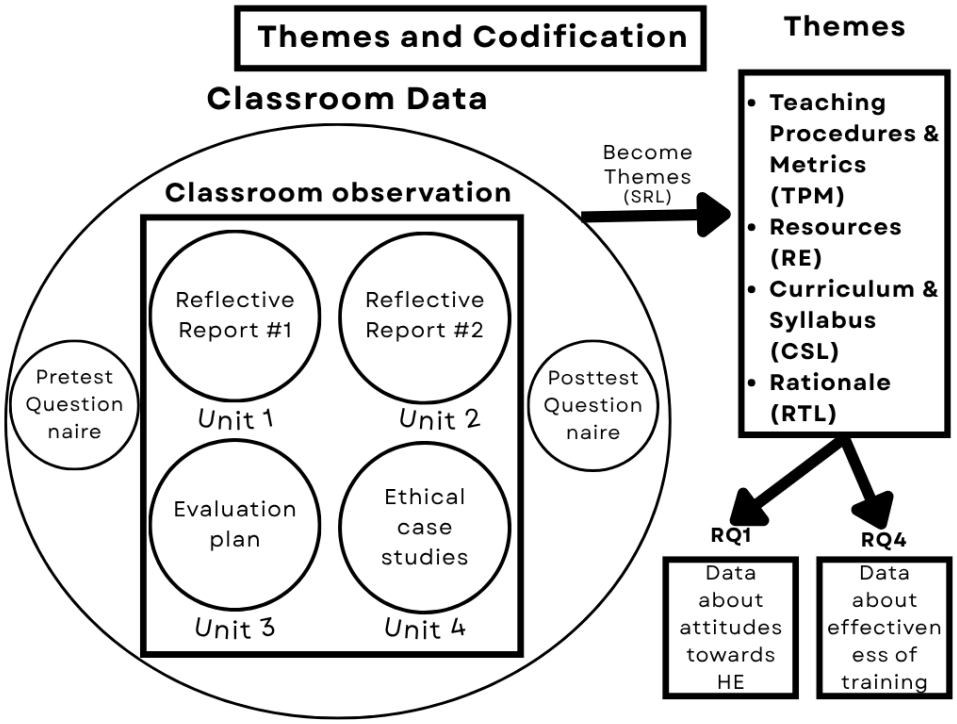


Figure 7.1: Figure detailing Phase 5 - Evaluation. It considers the process of coding themes from classroom data. The classroom observation was performed across four units, to collect data from reflective reports, evaluation plans, and case studies. The data informs the thematic categories, resulting into themes from the SRL, which address RQ1 and RQ4.

7.1 Themes and Nvivo Coding

Themes were coded into TPM, RE, CSL, RTL, following the themes from the SRL (Chapter 3, Subsection 3.4). The process can be described in Figure 7.1, showing how the students' assignments, reflections, and comments were structured into themes. The subthemes, as seen in Table 7.1, were categorised by the relevance of the comments in the assignments according to the content. Not all themes or subthemes were present in all units, they will be described as they occurred in each unit. Each theme and subtheme will be discussed in the order of the table, except for CSL in Unit 1 which will be discussed first (Subsection 7.1.1.1) and in Unit 4 which will be discussed last (Subsection 7.1.4.3) as they refer to the beginning and the end of the course, respectively. Where relevant, I will format part of the students' quotes in bold to highlight pertinent words to aid in the analysis.

The pretest questionnaire (Appendix J) recorded students' baseline understand-

Theme	Subthemes
RTL – Understanding and interpretations of MTQA	Ethical Considerations; Evaluation Frameworks & Criteria; Concept of Quality; Metrics; UX/UI; Human-Centred & Critical Perspectives; Reflections on MT use; Caution with MT use; Purpose of MT use; Transferability to other NLP tasks
RE – Uses and comments about resources	Standardised Norms; Platforms
TPM – Evaluating classroom aspects	Teaching delivery; Difficulty levels; Contextual factors
CSL – Considerations over the content	Expectations over Content; Pedagogical Aspects; Transferability of Skills

Table 7.1: Categorisation of themes and their corresponding subthemes identified in the analysis.

ing of topics taught in two aspects: i) their attitudes towards HE and AEMs and; ii) pedagogical aspects, such as expectations of the course. Each aspect is investigated **chronologically** to respond to RQ1, RQ3 and RQ4. For each unit (from 1 to 4), the themes TPM, RE, CSL and RTL were used to build an analysis of how they respond to the RQs in those respective units. To avoid duplication in the theming, a matrix coding query was conducted to identify co-occurrences between themes. The resulting matrix revealed that 3.9% of all possible unique theme pairs overlapped. The distribution of subthemes across units can be seen in Figure 7.2.

7.1.1 Unit 1

The heat map of the subthemes in Unit 1 can be seen in Figure 7.3 and the occurrences of each present theme can be observed in Table 7.2. For a full list of all subthemes, including those not present in this unit, see Appendix K.

7.1.1.1 CSL - Subthemes

To discuss this theme and its subthemes, a selection of questions provided the data from which the themes emerged, and only those cited will be focused on. For example, Q1 through Q3 were used to organise students' backgrounds in groups (Appendix G), while Q4 was effectively used to examine students' intentions with

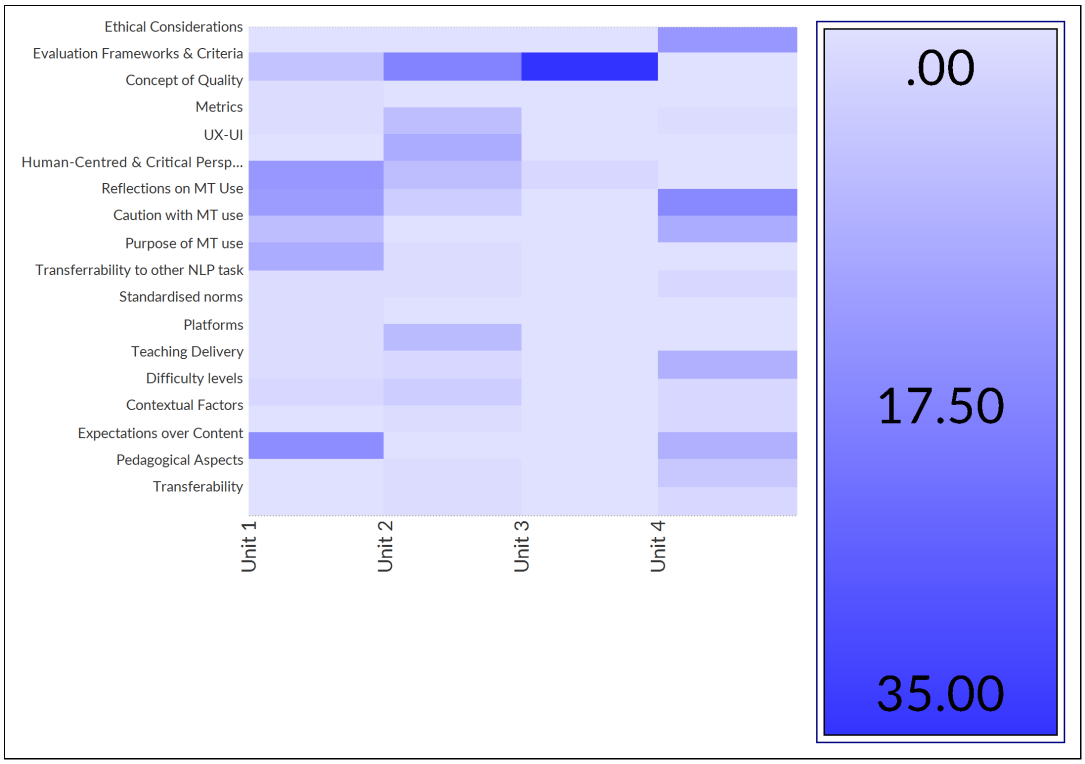


Figure 7.2: Heatmap illustrating the distribution of subthemes identified from TPM, RE, CSL, and RTL across four units. The intensity of the shading reflects the frequency of each theme's occurrence, with the highest observed frequency being 35 instances for a single theme.

the training, whether they wanted to learn about MTQA in general or with a specific focus:

- Participants who wished to learn more about HE - P4S4 (TT)
- Participants who wished to learn more about AEM - P4S11 (TT),
- Participants who wished to learn more about both HE and AEMs - P4S5 (NLP), P4S6 (TT)
- Participants who wanted a general understanding of MTQA - P4S1, P4S3, P4S8, P4S9, P4S15, P4S19 (TT); P4S2, P4S7, P4S10, P4S12, P4S13, P4S16, P4S17, P4S18 (NLP)

This data from the pre-test questionnaire shows that students had some prior knowledge. For example, Q11 (Appendix J) revealed that some students could define AEMs. P4S2 stated that AEMs aim to identify how well the MT output

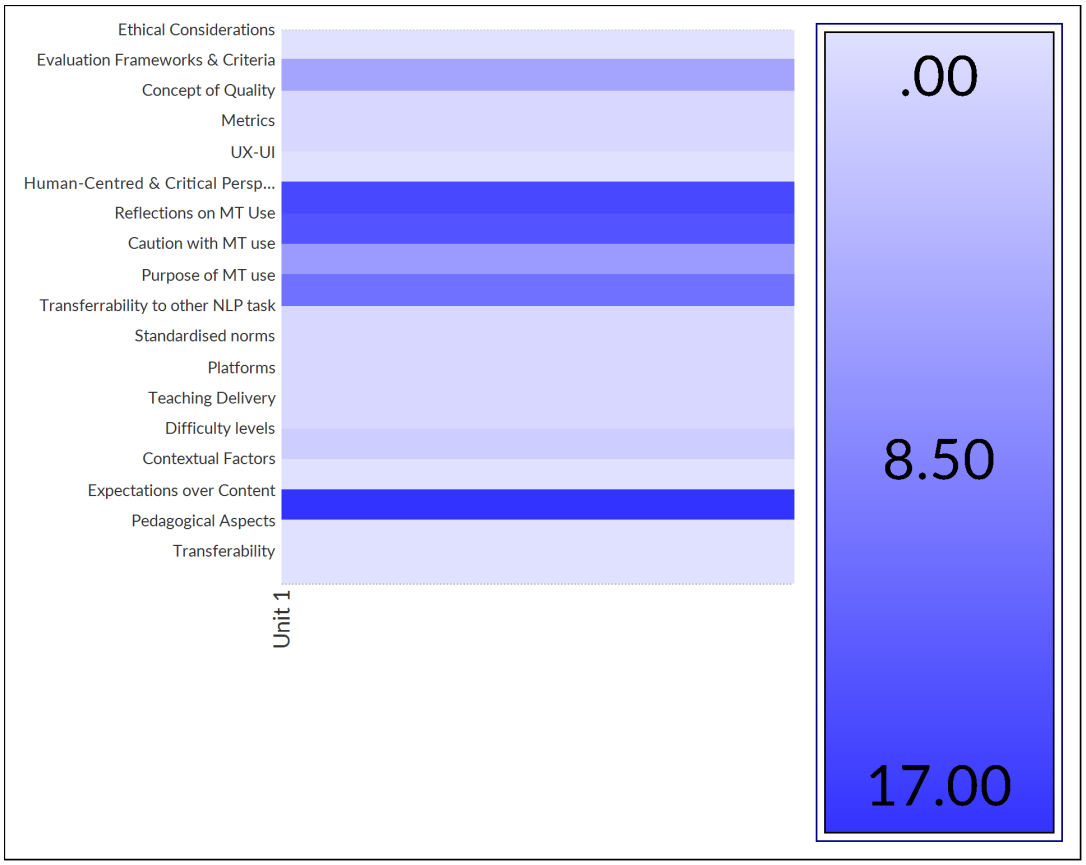


Figure 7.3: Heat map illustrating the distribution of subthemes identified from TPM, RE, CSL, and RTL across Unit 1. The intensity of the shading reflects the frequency of each theme's occurrence, with the highest observed frequency being 17 instances for a single theme.

correlates with a reference translation; P4S4, P4S6, P4S9, P4S11, P4S15 mentioned that AEMs are metrics that count similarity of n-grams; P4S5, P4S8 and P4S10, P4S12 described evaluation as a general diagnostic of the system. On the other hand, in Q12 (Appendix J), P4S2 and P4S15 mention HE as a process of evaluating MT output with translators using predefined criteria, while P4S13 mentioned that HE is a time consuming process that can introduce human bias and errors. Not many students responded to Q12 compared to Q11, which led me to believe that they were more familiar with HE than AEM.

7.1.1.2 RTL - Subthemes

In Unit 1, Lesson 2, students manifested more of their thoughts, reflecting while participating in the learning activities. As mentioned in Chapter 6 (subsection

Subtheme	Occurrence	Percentage
(CSL) Expectations over Content	17	21.79%
(RTL)Transferability to other NLP tasks	1	1.28%
(RTL) Purpose of MT use	11	14.10%
(RTL) Caution with MT use	7	8.97%
(RTL) Reflections on MT use	14	17.95%
(RTL) Human-Centred & Critical Perspectives	15	19.23%
(RTL) Metrics	1	1.28%
(RTL) Concept of Quality	1	1.28%
(RTL) Evaluation Frameworks & Criteria	6	7.69%
(RE) Standardised norms	1	1.28%
(RE) Platforms	1	1.28%
(TPM) Teaching Delivery	1	1.28%
(TPM) Difficulty levels	2	2.56%

Table 7.2: Occurrences and percentages of subthemes coded in Unit 1.

6.3.2), the students realised how difficult it is to perform HE. Whether this was due to lack of expertise or structured guidance (through guidelines), it is possible that it was from this moment in the lesson that students quickly realised that MT is a task that still requires ongoing work and expertise. This can be seen from Reflective Report #1, which led to the subtheme of **Reflections on MT use**:

- NLP Participants expressed limited knowledge of how translation works and what evaluation entails. (P4S5, P4S10, P4S16)
- NLP Participant thought MT was a solved problem, but changed their mind after Lesson 2. (P4S5)
- NLP Participants were very concerned about the quality of MT for low-resourced languages. (P4S5, P4S16)
- TT Participants expressed that working with HE through different perspectives (such as of an end user) made them look at MT in a different way,

not focusing on linguistic criteria as translators would, being more forgiving towards MT. (P4S1, P4S3, P4S15)

- TT Participants expressed no change in how they use MT, as they were already aware of the need for human intervention. (P4S9, P4S11)

Participants also thought about what a change in the end-user's perspective entailed in **Reflections on MT use**.

- TT participant mentioned it was difficult to move from a "the text must be perfect" mindset into a "as long as we get the message, it can be enough". (P4S1)

- TT participants reported less stress when using MT as an end-user, as they only had to gist. (P4S6, P4S8 P4S11, P4S15)

- TT participant reflects about perishability, saying that "serious scenarios" require a professional and post-editing to a high extent, but for casual needs, the MT output can be acceptable. (P4S4)

- NLP participants also reported less anxiety and less focus on details in comparison to the previous exercise. (P4S5)

The reflections increasingly included more details as students recounted their experiences of the progress of the lesson, focusing on **Caution with MT use**, as can be seen in their opinions:

- NLP Participants became cautious after the lab after realising MT is not perfect, when noticing the output can be unreliable due to grammatical mistakes or meaning not being carried through. (P4S5, P4S7, P4S10)

- TT Participants mention caution with MT as the output always requires review and refinement. (P4S11, P4S15)

- TT Participant urged risk assessment to be performed as MT brings risks when used inadequately. (P4S15)

- TT Participant became more aware of the risks after seeing MT systems' performance with low resource languages. (P4S8)

Reflections also led to students thinking about the human when interacting with MT with **Human-Centred & Critical Perspectives**:

- TT participants acknowledge how important it is to edit the output of MT with linguistic criteria and expertise on terminology and cultural knowledge, and acknowledge how much effort is put into this process. (P4S4, P4S6, P4S11, P4S15)

- TT participant also mentions that despite MT acting as a "word advisor" when checking the output, they still believe that the sentences are produced by the translator, not by the MT system. (P4S3)

- NLP participant also sees the value of expert humans interacting with MT, mentioning how they gained respect for translators. (P4S5)

Reflections on the **Metrics** showed the following:

- NLP participants were surprised by the disagreements in fluency using IAA measures, thinking there would be more agreement among peers. (P4S12, P4s16)

- NLP participant thought subjectivity was difficult to explain without a translation skillset. (P4S5)

- TT participants were not surprised that IAA measures were not always in agreement when comparing fluency scores. (P4S3, P4S6)

These comments have demonstrated that, when interacting with MTQA tasks, both NLP and TT participants have reflected on their practice, whether from a professional perspective or as regular users of MT systems. The practical components coupled with a reflective assignment have made NLP students realise the complexity of HE, especially in contexts involving low-resourced languages, even making

them question the idea that MT is a "solved" problem. P4S5, for example, stated "Initially, while knowing it's not true, I guess **I thought that MT is a 'solved" problem** before today's lab". TT students, while aware of the need for human intervention, appreciated the nuance of gisting versus professional use and started to think about risks and scenarios in which it is acceptable or not to use MT in this way. P4S1 (TT), for example, stated "I will start **to look differently at MT** for certain" in their report. P4S15 (TT) "**I am more aware of the end user now**, not just the translator as the system should work as well for both".

I believe that this initial, practical contact has shown evidence that something had changed in terms of attitudes for students. P4S16 (NLP) stated "After considering from both perspectives **I became much more critical** than when I was only engaged in one aspect of the assessment", mentioned how learning about HE changed their perspective. Having comments showing appreciation and respect for translators is a way of admitting how complex the task of responsibly using and evaluating MT systems is. Furthermore, while subjectivity has been pointed out as a weakness or difficulty of HE, at this point the students had started mentioning instructions or guidelines to decrease that variability in decisions, demonstrating that they had started thinking about solutions since Lesson 2.

7.1.1.3 RE - Subthemes

NLP participant P4S2 stated that the tasks on Google Sheets were useful.

7.1.1.4 TPM - Subthemes

In terms of feedback, NLP participant P4S2 reported satisfaction with the first week of lessons, saying that hands-on experience with the HE metrics was useful. P4S12 compared the task that involves comprehensibility with another NLP task, sentiment analysis, and revealed that it made the task more straightforward. TT participants also discussed difficulty levels. P4S7 found that the lab was quite subjective and not that challenging, while P4S9 felt unsure if they did the right job as they did not

feel confident in their command of English.

7.1.2 Unit 2

The heat map of the subthemes in Unit 2 can be seen in Figure 7.4 and the occurrences of each present theme can be observed in Table 7.3.

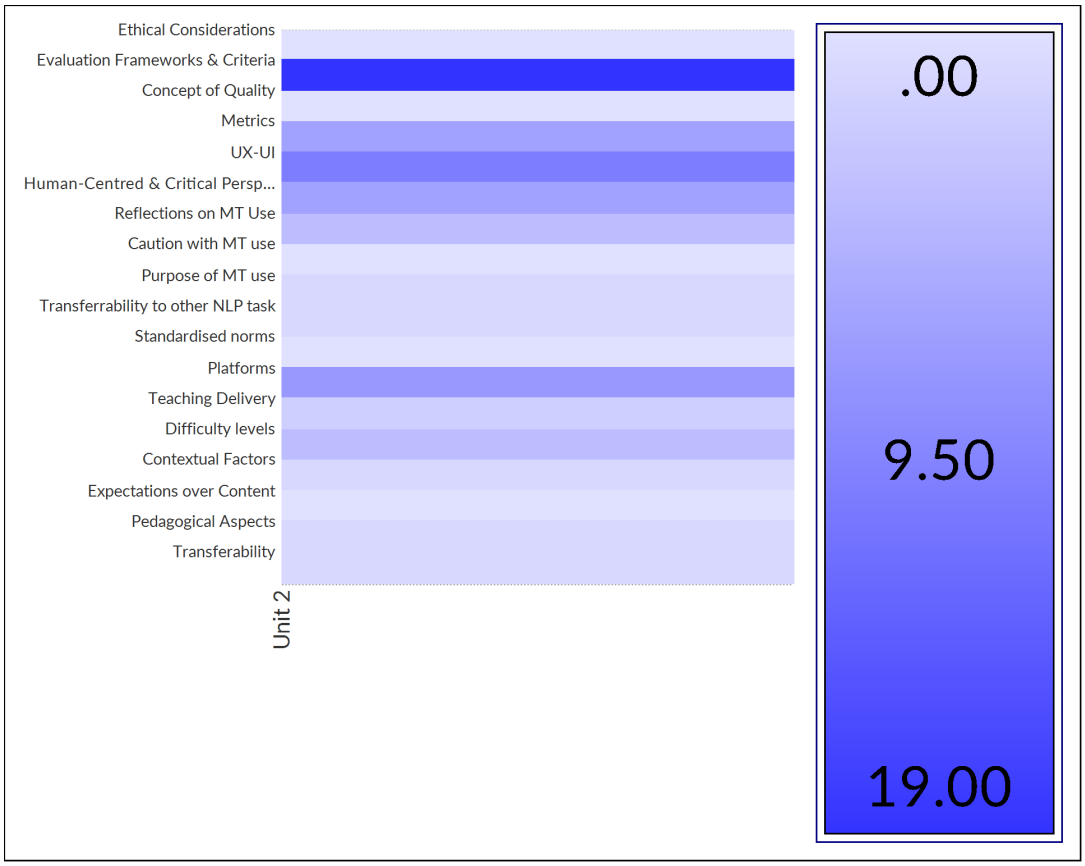


Figure 7.4: Heat map illustrating the distribution of subthemes identified from TPM, RE, CSL, and RTL across Unit 2. The intensity of the shading reflects the frequency of each theme's occurrence, with the highest observed frequency being 19 instances for a single subtheme.

Regarding Unit 2, much like Unit 1, students' thoughts and opinions were found in the assignments of this lesson (see Subsection 6.3.4), in Reflective assignment #2.

7.1.2.1 RTL - Subthemes

This lesson had a greater emphasis on document-level metrics (Tasks 2, 3, 5) in comparison to the previous one, which focused on sentence-level metrics. This led

Subtheme	Occurrence	Percentage
(RTL) Transferability to other NLP tasks	1	1.49%
(RTL) Purpose of MT use	1	1.49%
(RTL) Reflections on MT use	4	5.97%
(RTL) Human-Centred & Critical Perspectives	7	10.45%
(RTL) UX/UI	11	16.42%
(RTL) Metrics	7	10.45%
(RTL) Evaluation Frameworks & Criteria	19	28.36%
(RE) Platforms	8	11.94%
(TPM) Teaching Delivery	2	2.99%
(TPM) Difficulty levels	4	5.97%
(TPM) Contextual Factors	1	1.49%
(CSL) Transferability of Skills	1	1.49%
(CSL) Pedagogical Aspects	1	1.49%

Table 7.3: Occurrences and percentages of subthemes coded in Unit 2.

to students expressing how they felt about learning new approaches in Reflective Report #2:

- NLP participants expressed great interest in BlonDe as a document-level metric. (P4S2, P4S10, P4S12, P4S18 and P4S19)
- TT Participant expressed interest in BlonDe as a document-level metric. (P4S4)
- NLP Participant expressed interest in using BERTScore as they liked the idea of using transformer models for metrics. (P4S16)
- NLP Participants agree that combining sentence-level AEMs with document-level AEMs is ideal. (P4S10, P4S2)

Evaluation Frameworks & Criteria involved more emphasis on a combination of HE tasks and IAA measurement:

- NLP participants were able to find the strengths of document-level approaches. (P4S2, P4S5 P4S16, P4S18)

- TT participants were also able to find the strengths of document-level approaches. (P4S3, P4S4, P4S9)

- NLP participants were able to make comments about the strengths and weaknesses of specific settings for Likert scales. (P4S2, P4S5, P4S10, P4S12, P4S16, P4S18)

- TT participants were able to make comments about the strengths and weaknesses of specific settings for Likert scales. (P4S3, P4S4, P4S9)

By examining the human aspects, some students focused on the experience of evaluating, that is, on **UX/UI** aspects:

- NLP participants criticised inconsistent UI aspects and expressed how a long evaluation made them exhausted and even strained their eyes. (P4S12, P4S18)

In the reflective reports, it can be further observed that the participants were willing to implement the metrics they learned in their practice. For instance, P4S2 (NLP participant) felt motivated by pre-trained metrics and experimenting with different IAA measures. P4S3, P4S4, and P4S9 (TT participants) mentioned that they would like to combine sentence-level approaches with document-level approaches.

7.1.2.2 RE - Subthemes

Because many Platforms were used in this lesson, the participants discussed the following:

- NLP participant was surprised that platforms can provide different experiences, whether by making a workflow faster, smoother or more frustrating. (P4S12)

- NLP participants found Jupyter Notebooks useful. (P4S2, P4S12)

- TT participants found Jupyter Notebooks useful. (P4S3, P4S4, P4S9)

Based on these comments, the shift to document-level approaches has made NLP students show enthusiasm for metrics such as BlonDe and for using resources such as Jupyter Notebooks. Students (both NLP and TT) also showed interest in document-level approaches to HE, revealing that the context has given them more information that changed their evaluations. From this point on, participants began to analyse specific aspects in HE design. For example, they have shown greater reflection about what Likert scale settings are more appropriate to use depending on the context. Or they have inspected certain features in the UI that can affect the UX of the evaluator, which has made students realise that fatigue or frustration can decrease the quality of the evaluation. P4S18 (NLP), for example, stated "I found myself **straining my eyes** sometimes because some texts were too small or the opacity was reduced, or maybe it was the colour", and P4S12 stated "The **poor interface design** diminished the **effectiveness and enjoyment** of the experience".

From this moment on, students have employed language that implies more proactivity in using the metrics and resources learned in the course, such as"I am **interested in** using BERTScore" (P4S5, NLP), "I **would like to combine** several AEMs" (P4S4, TT), "I **would love to implement** them to my practice of MT evaluation" (P4S9, TT). At some points they showed uncertainty, such as P4S18 (NLP), who mentioned "Even though the document span was good and I could fairly evaluate some, **I wasn't quite confident doing so**". This comment shows the need for more structured instructions or guidelines for a real-scenario HE task, or at least more scaffolding from the teacher to explain the purpose of the exercise.

7.1.2.3 TPM - Subthemes

As mentioned in the difficulties when teaching Lesson 4 (Subsection 6.3.4), students reported some difficulty with the tasks. P4S2 (NLP) stated " Rating scales between 1 to 100 with interval units were very difficult", reporting the customisation made the experience harder to understand. P4S4 (TT) mentioned "I only tried one type of measure for IAA, it is a bit difficult to understand because I am not very profi-

cient in Python", referring to the Jupyter Notebook, which may not have been as straightforward for students without a coding background. However, some positive feedback was also offered. P4S10 (NLP) stated "**Your explanation** of the different scales and settings for evaluation types **was eye-opening**. Experiencing the differences made the purpose even clearer.", and P4S2 (NLP) "Overall, the lab was **a good demonstration** of using different AEMs and HE metrics", which shows that despite difficulty, some students mentioned positive aspects.

7.1.2.4 CSL - Subthemes

P4S5 (NLP) stated that coming from a non-MT background, they intend to test evaluation metrics in generalised text generation tasks, under the subtheme of Transferability of Skills. However, as a Pedagogical Aspect critique, P4S5 also stated that the structure of the lesson led to time constraints, so they had difficulty in looking at all tasks comprehensively.

7.1.3 Unit 3

The heat map of the subthemes in Unit 3 can be seen in Figure 7.5 and the occurrences of each present theme can be observed in Table 7.4.

Subtheme	**Occurrence**	**Percentage**
(RTL) Human-Centred & Critical Perspectives	2	5.41%
(RTL) Evaluation Frameworks & Criteria	35	94.59%

Table 7.4: Occurrences and percentages of subthemes coded in Unit 3.

In Unit 3, specifically Lesson 6 (Subsection 6.3.6), the students applied their knowledge of the metrics learned in Units 1 and 2.

7.1.3.1 RTL - Subthemes

All the **Evaluation Frameworks & Criteria** subthemes were coded as they were described in Lesson 6. They were complemented by the data acquired with classroom

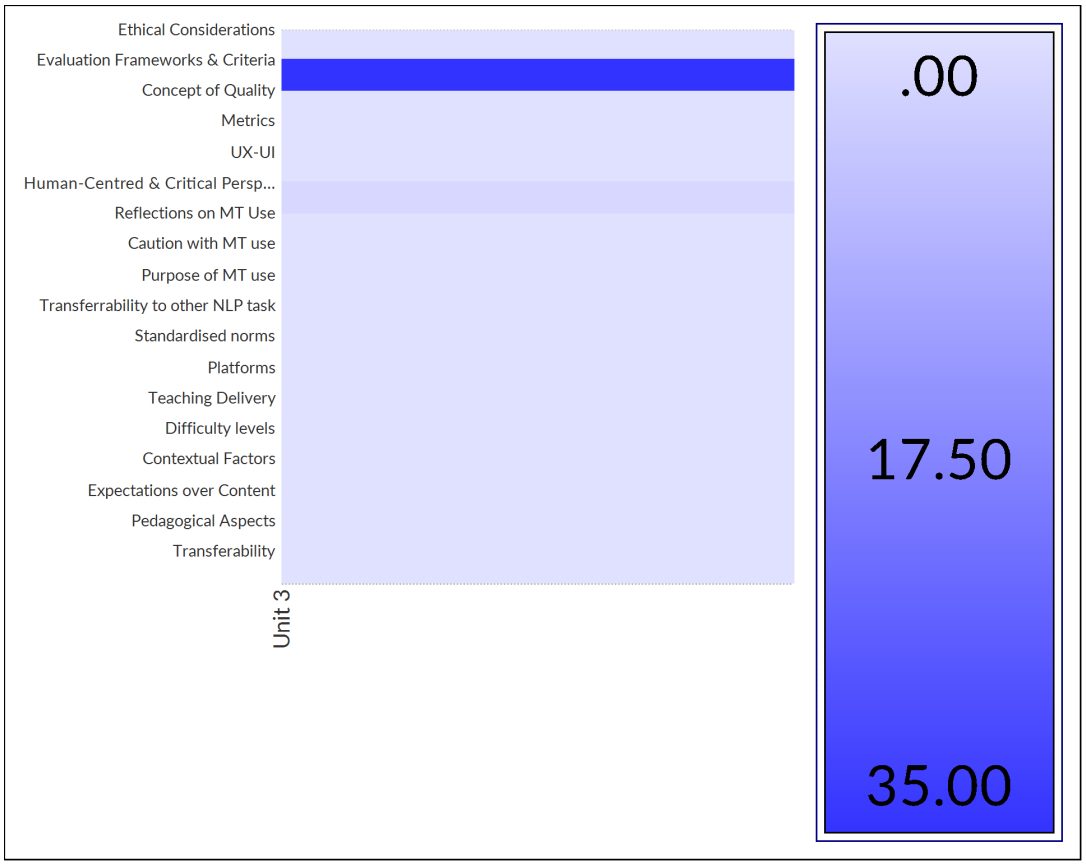


Figure 7.5: Heatmap illustrating the distribution of subthemes identified from TPM, RE, CSL, and RTL across Unit 3. The intensity of the shading reflects the frequency of each theme's occurrence, with the highest observed frequency being 35 instances for a single subtheme.

observation to support the data from this subtheme.

- Group 1 (NLP - P4S12, P4S16; TT - P4S1, P4S15) had issues with the scope of the project.
- Group 1 (NLP - P4S12, P4S16; TT - P4S1, P4S15) discussed in a very opinionated way the right proportion of test sets for HE and AEMs.
- Group 1 (NLP - P4S12, P4S16; TT - P4S1, P4S15) discussed features of informal language.
- Group 2 (NLP - P4S13, P4S10, P4S17, P4S18) wanted to mention the course has been useful.
- Group 2 (NLP - P4S13, P4S10, P4S17, P4S18) asked for clarification about perishability.

- Group 2 (NLP - P4S13, P4S10, P4S17, P4S18) were using an LLM to retrieve lists of datasets.
- Group 1 (NLP - P4S12, P4S16; TT - P4S1, P4S15) and Group 2 (NLP - P4S13, P4S10, P4S17, P4S18) both discussed the ethical implications of testing different MT systems considering the type of data their briefs were handling.
- Group 1 (NLP - P4S12, P4S16; TT - P4S1, P4S15) mentioned they wanted to test an LLM performing a MT task, namely Llama, as their criterion for adoption was its open source status.
- Group 3 (NLP - P4S7; TT - P4S4, P4S9, P4S8) were using the LLM Perplexity with ChatGPT to brainstorm ideas and gather resources.
- Group 1 (NLP - P4S12, P4S16; TT - P4S1, P4S15) were discussing what HE methodology would be appropriate, whether they focused on sentences, context-span or document-level).
- Group 3 (NLP - P4S7; TT - P4S4, P4S9, P4S8) mentioned they would build a test suite by combining sentiment analysis and emojis to complement the MT task.

Group 2 (NLP - P4S13, P4S10, P4S17, P4S18) reported to me that "The knowledge of this course helped us in our course work, **we feel more confident** in doing activities related to HE metrics and AEM metrics". Group 1 (NLP - P4S12, P4S16; TT - P4S1, P4S15) expressed that they were having trouble deciding the scope of the evaluation, as the options could be endless. I suggested for them to decide if they were working on a tight project (low budget, little time) or a more generous project (huge budget, plentiful time) to ground their evaluation on more solid terms. Several elements of the discussions showed that the students felt comfortable making decisions and defending them among their peers. For example, Group 2 had a mix of NLP and TT participants, so it was possible to notice the contrast of points of view that had different focuses. This activity also showed that the groups were willing

to combine LLMs in the process, either to structure information or to suggest the use of an LLM to perform a MT task. This shows their willingness and curiosity to experiment with existing MTQA approaches with LLMs, reminiscent of what Freire (1996) mentions of a critical curiosity for what students are learning.

7.1.4 Unit 4

The heat map of the subthemes in Unit 4 can be seen in Figure 7.6 and the occurrences of each present theme can be observed in Table 7.5.

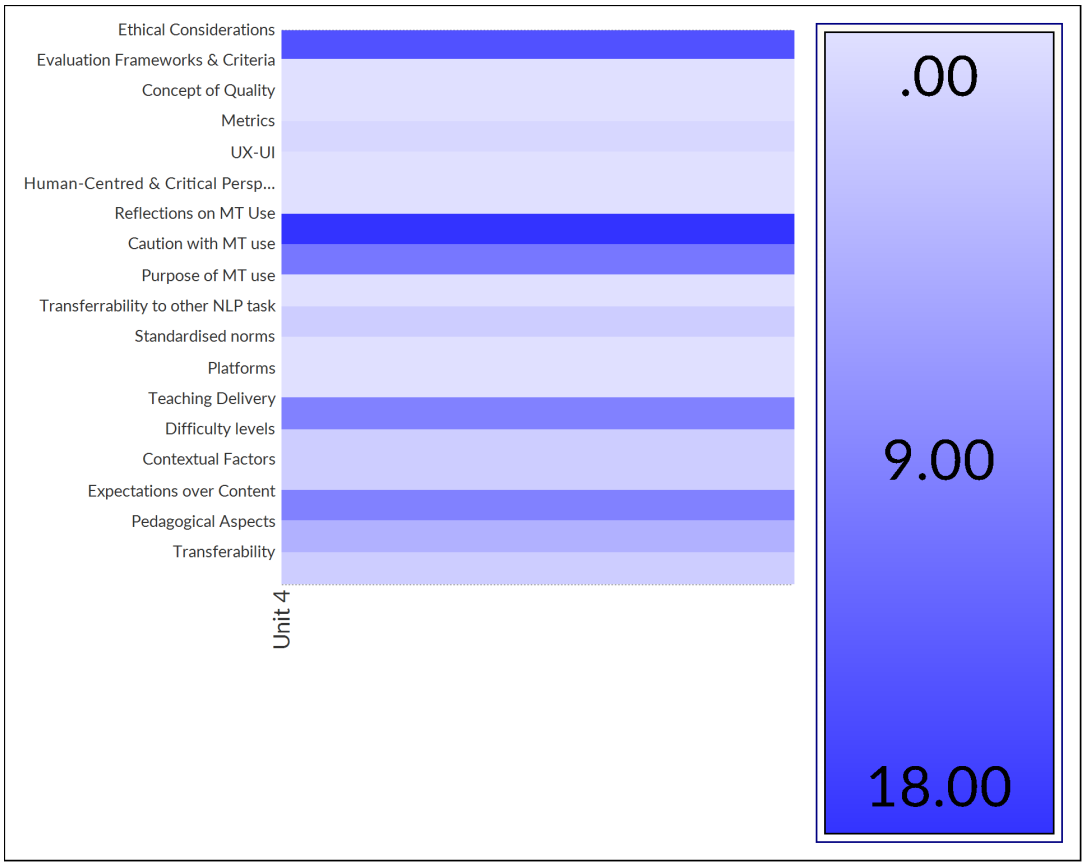


Figure 7.6: Heatmap illustrating the distribution of subthemes identified from TPM, RE, CSL, and RTL across Unit 4. The intensity of the shading reflects the frequency of each theme's occurrence, with the highest observed frequency being 18 instances for a single subtheme.

The themes of this unit focused on ethics. All data were acquired in Lesson 8 (Subsection 6.3.8). Following the notes described in the subsection, related to **Ethical Considerations**:

- NLP participants were sceptical of learning activities that contained hyperbolic

Subtheme	Occurrence	Percentage
(RTL) Transferability to other NLP tasks	2	2.56%
(RTL) Caution with MT use	11	14.10%
(RTL) Reflections on MT use	18	23.08%
(RTL) Metrics	1	1.28%
(RTL) Ethical Considerations	15	19.23%
(TPM) Teaching Delivery	10	12.82%
(TPM) Difficulty levels	2	2.56%
(TPM) Contextual Factors	2	2.56%
(CSL) Transferability of skills	2	2.56%
(CSL) Pedagogical Aspects	5	6.41%
(CSL) Expectations over Content	10	12.82%

Table 7.5: Occurrences and percentages of subthemes coded in Unit 4.

claims about new MT/LLM systems. (P4S5, P4S8, P4S16)

- NLP participants reported wary of giving away their personal data to tech companies, as it could be used to train LLMs. (P4S5)
- NLP participant recalled from one of their core modules that transparency is essential. (P4S8)
- NLP participant stated that without the training, they would believe hyperbolic claims about AI-driven translation. (P4S16)

From this lesson, the first task involving the case studies was the task that achieved the most engagement from students, as over the lesson, their responses entailed more describing papers rather than making their own comments. Afterwards, the students expressed themselves more openly in the posttest questionnaire (available in Appendix H) responses, from which most of the subthemes were extracted in this unit.

7.1.4.1 RTL - Subthemes

When asked what technical aspects are essential for MT (Q2), students had **Reflections on MT use**:

- NLP participants said that knowledge of both HE metrics and AEMs is essential, especially to reach a reproducible and comprehensive evaluation. (P4S2, P4S5, P4S7, P4S16)
- NLP participant stated that understanding the purpose of the system and how it will be used is necessary to decide the best evaluation approaches. (P4S12)
- NLP participant stated that considering some amount of HE is essential, to figure out where it is working or not rather than focusing on numerical scores. (P4S13)
- NLP participant stated that having an evaluation planned is important to cover all aspects of MTQA. (P4S18)
- TT participant stated that programming knowledge combined with language and cultural sensitivity, is important. (P4S4)
- TT participant mentioned that measuring energy consumption is essential. (P4S8)
- TT participant said that knowledge of both HE metrics and AEMs is essential, especially to reach a reproducible and comprehensive evaluation. (P4S9)
- TT participant mentioned that rigour in evaluation is important, especially to reach statistical significance in results. (P4S19)

Regarding **Ethical Considerations** on Q3:

- NLP participant highlighted the importance of being ethical about how data is acquired, used and processed. (P4S2, P4S12)
- NLP participant mentioned that the performance of a system should not be more important than the impacts on the planet, and all emissions should be reported. (P4S5, P4S16)
- NLP participant mentioned that handling biases is essential. (P4S5)

- TT participant stated that human-centric and environmentally friendly MTQA is essential. (P4S4, P4S13)

- TT participant mentioned that it is not healthy for the development of MT systems to replace people. (P4S4)

- TT participants defended the use of qualified professionals and their intellectual property, mainly translators. (P4S1, P4S9)

- TT participants stated that transparency is essential, as hype over systems can be dangerous, especially if data was acquired in an unethical way. (P4S8)

- NLP participant stated that responsibility and valuing translators are important factors in MTQA. (P4S10)

- NLP participant stated that the TBL is an essential ethical aspect in MTQA. (P4S18)

- TT participant expressed that reliability of results is important, as results must be shown in a truthful way, whether good or bad. (P4S19)

Focusing on **Caution on MT use**, participants responded in Q4:

- NLP participant stated that results could be unreliable and not reflect the true performance of the MT system, creating potentially risky, biased and environmentally damaging MT systems. (P4S2, P4S5)

- NLP participant stated that improper evaluation can lead to MT systems that deliver missing, wrong or biased information. (P4S13, P4S18)

- NLP participant stated that if results are not reliable or reproducible, progress would be slowed. (P4S16, P4S18)

- TT participant stated that all claims about quality could be subjective and difficult to tell their validity. (P4S1, P4S8)

- TT participant mentioned that it is not good for the future of MT, as it could feed the misconception that MT is good enough to replace humans with systems that do not have acceptable quality. (P4S4, P4S9)

- TT participant said that not only would results be unreliable if not transparent, but that they would cost a lot of energy consumption for no reason. (P4S19)

In general, NLP participants prioritised methodology, reproducibility, and use cases for systems. In contrast, TT participants considered other aspects of MTQA, such as programming skills, cultural sensitivity, and more attention to sustainability. However, both groups agreed that comprehensive evaluation must combine HE with AEMs, suggesting that perhaps they are aware of the strengths and weaknesses of each approach, leading to the decision to combine both.

Ethical concerns also had different framings. NLP participants tended to focus more on data and impacts on the environment. However, TT participants had a greater focus on working conditions and respect for the labour of translators, which makes sense considering their background.

7.1.4.2 TPM - Subthemes

In terms of Teaching Delivery, the following comments were made:

- NLP participants stated that they enjoyed the theoretical and practical components. (P4S2, P4S12

- TT participants stated that they enjoyed the theoretical and practical components. (P4S1, P4S19)

- TT participant stated that it was a good decision to mix students' backgrounds, as they could hear other perspectives. (P4S9)

- NLP participants stated that they appreciated the labs backed by academic literature. (P4S2, P4S5)

- NLP participants stated that they enjoyed interactive tasks and group tasks. (P4S2, P4S18)

- NLP participant stated that they enjoyed tasks being broken down into steps. (P4S18)

When observing this subtheme, my teaching was described as "enthusiastic" (P4S5, NLP), "engaging" (P4S5, P4S16 NLP), "well delivered" (P4S5, NLP). The classes were complimented for combining the literature to support MTQA approaches and metrics, as stated by P4S2 (NLP) as "well researched and backed by the academic literature" or by P4S5 (NLP) that the course "highlighted a world of papers in the field that **could be of use to me and my future research**". The approach of using tasks informing both theory and practice was commended by participants, such as P4S18 (NLP) "This course used a breakdown of steps to **theoretically and practically inform and teach us**". For students, group work was also highlighted as positive, as stated by P4S2 (NLP) "The group tasks were really very **impressive** and **helpful**. **I feel very included and involved** in each activity", P4S9 (TT) "**Group work really helped me** understand the lectures", P4S18 (NLP) "**I like the practical aspects** and the group tasks."

In **Difficulty levels**, P4S4 (TT) expressed that not having a background in NLP made things more difficult, but that difficulty was used for motivation: "I actually want to learn more details about programming things". **Contextual factors** were also reported, with P4S1 (TT) stating that the course hours were a bit late, which could have impacted their learning. P4S18 (NLP) stated that not being able to attend classes for personal reasons affected their experience.

7.1.4.3 CSL - Subthemes

For the **Pedagogical aspects**:

- NLP participants stated that the duration of the course was short but effective. (P4S2)

- TT participants stated that the duration of the course was short but effective. (P4S4)
- TT participant stated that the structure of the course was ideal to have a comprehensive understanding of MTQA. (P4S4)
- TT participant stated that they did not have enough time to read the papers suggested in the classes. (P4S4)

When expressing **Expectations over Content**:

- NLP participants felt that more live interactions with NMT or LLMs should have been implemented, particularly for testing. (P4S2)
- NLP participants stated that the training was successful in introducing them to evaluation metrics and the field overall. (P4S5)
- TT participants had more contact with tools than expected, but realised that the time spent on papers was just as useful (P4S8)
- NLP participants said the course exceeded expectations due to the combination of studying the design and implementation of MTQA. (P4S10)
- NLP participants mentioned that the course met their expectations as they were unaware of the importance of HE. (P4S13)
- NLP participant mentioned that there could have more attention dedicated towards explaining how to read the results of AEMs. (P4S13)

Some comments were made about how the knowledge of the course could be transferred to the participants' programmes. P4S8 (TT) stated that "This training met my expectations, **I used the knowledge from this training in my core lectures**. And they will definitely affect my assignment grade". P4S9 (TT) expressed that the course was confusing at first but was easier to follow as they studied their "Translation Technology" module, which led them to state "I never considered the

ethical aspect of this topic, thanks to the course I got to know this part, such as sustainability".

7.2 What can be taught to NLP students to encourage them to carry out HE?

At the beginning of the course, although my main focus was fostering the use of HE, I chose to also measure students' confidence with AEMs to see if results would influence HE in some way. As collected from Q9 and Q10 from Appendix G to serve as baseline scores, students started with confidence in HE with an average of 2.11 and confidence in AEM with an average of 2.37 (see Appendix I). These scores are relatively at the lower end of the scale (which ranged from 1 to 5), implying that they did not know much about working with either type of metric. Drawing from the results of the pre-test questionnaire (see Section 7.1.1), at first the responses varied as to how students wanted to learn more about HE, AEMs or to have a general understanding of MTQA. I followed the Freirean approach that supports educators who know the interests or prior knowledge of students (Freire, 1996). Knowing the expectations and skills sets of the students, I could teach the NLP students to be aware of the strengths and weaknesses of each evaluation approach. This data from the pretest questionnaire served as a baseline to compare with, starting from the student feedback and reflections collected at the end of Unit 1, organised in Table 7.6.

7.2.1 Unit 1

At the end of this unit, I noticed that the students had expressed that they changed their minds about MT and how to evaluate it as I read the data extracted from Reflective Report #1, which was completed after the lab in Lesson 2. There is evidence of participants rethinking their position that MT is a solved task (P4S5, P4S16), particularly with the examples from low-resource languages (P4S5). The reflection

Unit 1 Reflections
• "Knowing about limitations of each metric in practice helps to think out of the box to use an efficient combination of these metrics for a specific task". (P4S2) • "I guess I thought that MT is a 'solved' problem before today's lab. Now I know this is not true". (P4S5) • "From a non translator background, this evaluation exercise gave me even more respect for human translators". (P4S5) • "Through these evaluation tasks, I gained a deeper understanding of the translation process". (P4S10) • "I had assumed MT was generally very accurate, but after seeing the differences during the class today, I will be more sceptical about MT output I see". (P4S16) • "After considering from both perspectives I became much more critical than when I was only engaged in one aspect of the assessment". (P4S16)

Table 7.6: NLP Participants reflections from Reflective Report #1 present in Lesson 2, Unit 1 where they demonstrated willingness to use HE or positive attitudes towards HE.

of P4S10 showed some evidence that studying MTQA helped the participant understand how translation works through the study of HE metrics, such as adequacy and fluency. P4S16 explicitly used the word "critical" to describe how they felt when learning about HE and AEM, compared to how they felt initially with only one type of metric. I believe that part of this "critical" perspective involves understanding what can be done combining the strengths and weaknesses of both types of metrics, which was perfectly described by P4S2. This matches the planning of the course (Chapter 5, Subsection 5.2.1), where the primary objective was that exposure to these tasks leads to a positive perception of HE and MTQA in general. This leads me to believe that the inclusion of metrics, both HE and AEM, was essential for students to perform the evaluation and gain initial confidence.

7.2.2 Unit 2

This unit has more examples of students writing comments in their reflective reports (Reflective Report #2) organised in Table 7.7. In this lesson, I observed that the students understood the strengths and weaknesses of document-level approaches and noticed how UX impacts evaluation, as noted in Subsection 7.1.2. More explicitly, I noticed students stating willingness to use certain types of metrics, such as "I would

Unit 2 Reflections
• "I would like to implement the pre-trained metrics for my own practice and also use IAA scores for analysing the agreement between human evaluators when doing HE". (P4S2) • "As I am new to this area, it was very interesting to note adding the context. It really enhanced my ability to evaluate the system". (P4S5) • "Your explanation of the different scales and settings for evaluation types was eye-opening". (P4S10) • "I think document-span evaluation was more helpful as you can focus on more precise issues in the translation". (P4S16) • "Document level was more helpful to me, it improved my understanding of the context". (P4S18)

Table 7.7: NLP Participants reflections from Reflective Report #2 present in Lesson 4, Unit 2 where they demonstrated willingness to use HE or positive attitudes towards HE.

like to implement pretrained metrics", as P4S2 stated in their report. I also noticed that the students verbally showed improvement, such as P4S5 and P4S10. The usefulness of the approaches was also reported; P4S16 and P4S18 have mentioned document-level approaches, involving either context span or the whole document, and how context plays an important role in evaluation. This Unit also corresponded to the expectation described in Chapter 5 (Subsection 5.3.1), where the tasks would encourage students to include UX as a factor in evaluation and gain more attention to alternate HE methodologies and the study of AEMs.

7.2.3 Unit 3

Unit 3 Reflections
• "The knowledge from this course helped us in our course work, we feel more confident doing activities related to HE metrics and AEMs". (P4S10, P4S13, P4S17, P4S18)

Table 7.8: NLP Participants reflections from the evaluation planning assignment present in Lesson 6, Unit 3 where they demonstrated willingness to use HE or positive attitudes towards HE

At this point, I reached the second half of the course, **Lesson 5**. I was uncertain as an educator whether this experience would empower students to be confident with HE or be willing to use new MTQA approaches. However, the statistical

results were surprising. As seen in Subsection 6.4.2, the mean of confidence in HE rose to 3.67, just as confidence in using AEM rose to 3.40. This shift suggested more confidence, although that would only be confirmed with the Bayesian LLM. Analysing the results, it was shown that this increase between the pre-test and mid-test was statistically significant. The reflective comments from Unit 1 and 2 are examples that students at this point were feeling more confident, suggesting a good correlation with the statistically significant rise in confidence. I hoped that teaching evaluation design and ethics would continue to keep participants motivated and confident, as these topics would increase the complexity by introducing various scenarios that warranted the combination of metrics and tools more cohesively.

I wanted students to have the opportunity to experiment, to be more autonomous in their decisions after acquiring the tools. My role involved support and creating conditions for students to explore and critically assess what makes sense to use within the scenarios presented to them, as recommended by Freire (1996): a good educator poses realistic situations for students to use the knowledge they co-constructed with their students. In the learning activity of Unit 3, the students suggested through their responses (available in Subsection 6.3.6) that not only they would use HE, but they would use both HE metrics and AEMs. The HE metrics mentioned were adequacy (Groups 1, 2, 3), fluency (Groups 1, 3), readability (Groups 1, 2), usability (Groups 2, 3) and acceptability (Group 3). The only explicit comment about the training influencing their required module in their M.Sc. was shown in Table 7.8, where Group 2 had indicated that the course was actively helping them in their core module.

7.2.4 Unit 4

In this Unit I noticed that the students looked tired based on their facial expressions, both in the lecture and in the laboratory session. I even asked for clarification, as I was a bit concerned. Freire (1996, p. 131) suggests that educators can ask for feedback, or they may state that they may not explain every topic in the most

Unit 4 Reflections
• "Knowledge of both HE and AEMs is essential for MTQA". (P4S2) • "It introduced me to the evaluation metrics, while highlighting a world of papers in the field that could be of use to me and my future research". (P4S5) • "Not only I gained insights into MT systems, but the design and implementation of evaluations made understanding MTQA easier and more vivid. I have a clear understanding of what each metric does, how they work together and the direction I should take when analysing specific results". (P4S10) • "The hands on practice with different techniques of evaluation provided a valuable technical experience, it will be quite helpful in future use". (P4S12) • "I never thought that HE plays a critical role and I did not know about AEMs shortcomings". (P4S13) • "From the classes I could attend, I understand the HE and automatic evaluation methods used in measuring the quality of MT systems". (P4S18)

Table 7.9: NLP Participants reflections from the post test questionnaire filled at the end of Lesson 8, Unit 4 where they demonstrated willingness to use HE or positive attitudes towards HE

ideal way, since educators do not know everything exactly and are not perfect. Fortunately, that was not the issue, as students clarified that such a complex topic requires time for them to process and think, which brought me relief. Freire (1996, p. 132) explains that this availability of dialogue, of checking in with students, is essential for true dialogic classes. Educators will not always have the answers to everything, but learning needs can be analysed and met through communication. Table 7.9 contained reflections that were written at the end of the last lesson, drawn from the post-test questionnaire. The participants were successful in understanding that both HE and AEMs are important, respecting the strengths and weaknesses of each approach, by the comments of P4S2, P4S10, P4S13, and P4S18. In addition, the students explicitly reported that the course content will be useful in their practice (P4S5, P4S12). This provides evidence that at the end of the course, students viewed HE positively, much like they verbalised the usefulness of what they had learned.

7.3 What can be drawn from TS when teaching MTQA for NLP postgraduate audiences?

In Chapter 4, a survey was conducted in Phase 1c (Subsection 4.1), which contained some partial responses on what elements from TS can be used to teach MTQA. The findings of Phase 1c consisted of an organised collection of responses, and some of these responses resonated from data collected in Units 1 through 4 in student feedback. The student feedback will be described unit by unit, and where relevant, data from the survey of Phase 1c will be linked to the students' reflections.

7.3.1 Unit 1

Unit 1 Reflections
• "I am also expecting to know more about the possible variations and problems of these metrics" (Referring to adequacy, fluency, comprehensibility and usability). (P4S2) • "This lab gave me insights as to why there isn't a stricter scale for all of these evaluation metrics". (P4S5) • "I like to have my definitions and formulae clearly displayed but I think in these cases it's not as easy to define in these cases". (P4S5) • "It was interesting to see that while our IAA moderately agreed, there were quite a few differences". (P4S5) • "I gained a deeper understanding of the translation process, including key criteria like adequacy, fluency". (P4S10) • "I believe clear guidelines are essential, as their absence affected my second task". (P4S12) • "When you discussed the importance of trying to ensure that the user does not form any bias in their judgements by randomising the order of which MT system was used for each sentence, it made me think about the unpredictable nature of human behaviour much more seriously". (P4S16)

Table 7.10: NLP participants' reflections from Unit 1 on how translation studies resources (e.g., evaluation criteria and judgement elicitation practices) influenced their understanding of NLP evaluation

At the end of this Unit, when analysing Reflective Report #1, students focused on HE, as observed in Table 7.10. The adoption of curiosity about or praise for metrics is evident. Adequacy, fluency, comprehensibility, and usability instigated curiosity (P4S2). More than curiosity, adequacy and fluency have been shown to be

essential and the reason why a participant understood more about the translation process (P4S10). Bias and subjectivity have been discussed as a factor that shaped understanding of HE weaknesses (P4S5, P4S16). For example, the IAA results have shown that the lack of guidelines can be confusing and can negatively affect the results (P4S12). Although there were no mentions of guidelines or specific instructions on how to perform HE, students still noticed their absence in tasks. In the data analysis of Phase 1c (subsection 4.1.2), in Q29 ("If you were to create a HE module in MTQA geared towards NLP students, what should be the main content?"), adequacy and fluency were among the metrics that were recommended the most between both NLP and TS educators, suggesting that perhaps these two metrics are essential in a HE course. As suggested in Q30 where educators provided more details, hands-on sessions would be necessary to incorporate these metrics. The students' comments in Unit 1 may have confirmed the effectiveness of this suggestion.

7.3.2 Unit 2

Analysing Reflective Report #2, this unit contained more comments where students reflected on the usefulness and operationalisation of the resources in their own practice, as noted in Table 7.11. Jupyter notebooks have been proven to be effective in teaching MTQA (Krüger, 2022), and students in this training have been proven to appreciate their use (P4S2, P4S12). Some reflections focused on scales (P4S10, P4S16), indicating the usefulness of analysing several elements of evaluation design. It also corresponds to the suggestions of the TS educators in Q29 of Phase 1c (subsection 4.1.2), prioritising the teaching of elements of the evaluation design. Other methodological elements that were praised and shown to be useful were document-level approaches (P4S5, P4S18), showing how much attention has been given to this particular facet of evaluation in MTQA (Castilho and Knowles, 2024), and highlighted by educators as a trend in Q16 ("What trends do you foresee in evaluation metrics that incorporate human judgement for MT systems?") of

Unit 2 Reflections
• "I like the idea of (Jupyter) notebooks where we only had to run different code segments to study and observe the impact of different measures". (P4S2) • "With sentence level AEMs, it was helpful to inspect each sentence to understand how each metric worked to calculate the scores". (P4S2) • "I think IAA can be very useful.. It helps with results to help others understand and think about the credibility of systems and processes". (P4S2) • "Context-span was very interesting to note how adding the context really enhanced my ability to evaluate the system". (P4S5) • "In terms of preferences, I'd select a 1-5 scale with labels for sentence-level evaluations". (P4S10) • "For document-level evaluation, I'd choose a 1-100 scale, as it allows for more detailed scoring across an entire text, offering a more nuanced perspective". (P4S10) • "The BlonDe Metric in a Jupyter Notebook was useful for understanding AEMs as effective tools". (P4S12) • "I think a scale of 1-5 is most suitable, it allows you to be somewhere in the middle. More importantly though I think is to stay consistent and always provide a marker for where on the scale you are". (P4S16) • "Document level (approaches) were more helpful to me. It improved my understanding of the context". (P4S18)

Table 7.11: NLP participants' reflections from Unit 2 on hands-on engagement with AEMs, scaling practices, and the effect of context in translation evaluation

Phase 1c, suggesting that it was a productive choice to include this paradigm of evaluation in the course. In Phase 1c, Q26 ("Provide comments or explanations for your previous answer here", which followed up on Q25, which asked educators to assess the importance of including evaluation metrics in the academic curriculum), one response in the NLP group was that AEMs are indispensable. That response echoes students' sentiments when learning about them (P4S2, P4S12).

7.3.3 Unit 3

Unit 3 did not involve many themes, only the responses from the assignments. The only data referring to the content of Unit 3 was found in the posttest questionnaire, where P4S16 responded to Q2 ("What technical aspects do you consider essential when developing the MT evaluation and why"), stating: "The evaluation plan is crucial and a key factor. Having a proper, well-detailed evaluation plan provides better understanding on how to evaluate". That is, the participant found great use in the pedagogical tool used in Lesson 6. Although the assignment itself was not

drawn from TS, many of the tasks contained concepts and questions drawn from TS.

7.3.4 Unit 4

Unit 4 Reflections
• "Reporting how data was collected and how much carbon was emitted. In order to be more sustainable we need to be more transparent about how much harm we are doing to the environment. We also need to be transparent about how data is being collected so we know if any harm is being done to people or if results are biased". (P4S5) • "Linguistic, Python coding basics, understanding of different automatic metrics and some statistics". (P4S10) • "Make sure to understand what the system is about and the purpose of it and how will it be used. Based on that information we need to decide what approach of evaluation should be used". (P4S12) • "Providing code and dataset for reproducibility. Comprehensive evaluation metrics". (P4S16) • "The Triple Bottom Line (People, Performance, Planet)". (P4S18)

Table 7.12: NLP participants' reflections from Unit 4 on ethical considerations, transparency, reproducibility, and responsible evaluation frameworks inspired by translation studies

Unit 4 elicited comments that reflected the importance of ethics and sustainability, compiled in Table 7.12. The comment from P4S5 is very reminiscent of Moorkens's (2022) comments on transparency with data use, ownership, and distribution, which were referred to in Lesson 7 (available in Chapter 5, Section 5.5.1). P4S12's comment seems to contain elements of the concept of perishability (Doherty et al., 2018) and evaluation fit for purpose (Way, 2018). Finally, the TBL is described as a framework that is desirable to promote ethics and sustainability.

7.4 How effective is the content proposed for HE in an MT evaluation module?

To verify the effectiveness of the training, it is worth inspecting the LO described in the development of the course (see Section 5.1) to check whether students had

accomplished the LOs either at their respective lessons or by the end of the training (to check LOs per Unit/Lesson, see Figure 5.1). Each table in this section contains the relevant subthemes identified in Section 7.1 for each respective LO.

7.4.1 Learning Outcome 1

LO1 – Develop and foster awareness of the role of HE in the development of MT systems		
Application of HE	**HE interacting with AEM**	**Evaluation Frameworks & Criteria**
HE taught how systems can be assessed and used for various reasons. (P4S2)	HE complements AEMs with in-depth insights. (P4S2, P4S5, P4S7, P4S10, P4S16)	Awareness of importance of Likert Scales. (P4S5, P4S10, P4S18)
Context is a unique strength of HE at document-level. (P4S4)		Reported scale labels as important in HE. (P4S12, P4S18)
HE can be understood through the analysis and practice of metrics. (P4S10)		Sliders are important in HE for good UX/UI. (P4S12)

Table 7.13: Participants' views on HE application, interaction with AEMs, and evaluation frameworks within MT development (LO1), based on themes from Section 7.1 extracted from reflective reports, assignments and questionnaires.

Evidence from assignments (Reflective Report #1 and #2) suggests that students achieved LO1 through both conceptual understanding of metrics and their practical application. In reflective reports and evaluation tasks, participants demonstrated an ability to articulate what HE is, why it matters, and how it interacts with other evaluation approaches. Furthermore, since HE metrics were presented from various perspectives, it can be seen that students noticed how they can be used for different purposes, as evidenced by P4S2's comment in Table 7.13. That also led to students reflecting on specific features of HE design as well. Likert scales, scale labels, and custom slider settings were debated by participants in their reflections (P4S5, P4S10, P4S12, P4S18), for example.

When highlighting the strengths of HE, participants often compared this type of metric with AEMs. According to Table 7.13, students compared and were eager to

combine both types of metrics, as they considered their strengths (P4S2, P4S5, P4S7, P4S10 and P4S16). Although NLP students are likely not able to identify errors or linguistic features produced by NMT (Loock, 2020), they certainly have gained greater insights on the strengths, limits, weaknesses of MT, and most importantly, to recognise the value of translators. At several points in Lessons 2 and 4, when performing HE tasks, NLP students constantly mentioned how they did not feel confident, with some expressing the need for translators. It is recommended to have translators as evaluators (Läubli et al., 2020), and because of that, I firmly believe that HE in MTQA is truly understood when one understands why translators must be included in the process at some point.

Combining the results of the thematic analysis with other evidence from the training brings evidence that students accomplished their understanding of LO1. For example, in the Vevox quiz (see Subsection 6.3.5) the majority of the students responded correctly to the questions related to HE (Questions 1, 4, 5 and 6). It also been disclosed throughout the training that subjectivity is the most reoccurring weakness. In a way, this corresponded to what was theorised in SRL regarding TPM studies (see Section 3.5.1.4), as subjective elements had to be reflected upon to recognise potential challenges. Students also pointed out weaknesses of specific methodologies, such as the exhaustion and the extensive time that document-level HE can cause. In Lesson 6 (Subsection 6.3.6) it was also possible to observe that students chose various HE metrics and combined them with AEMs in Tasks 3a and 3b. Task 3c in this lesson showed that every group chose translators. Task 3d also demonstrated that students included some sort of Likert scale to be included. The results of these assignments lead me to believe that students are not only aware of HE, but are also willing to implement it into their workflows.

7.4.2 Learning Outcome 2

This LO was embedded more specifically in Unit 2, particularly in lessons and labs introducing sentence-level and document-level AEMs (Lessons 3 and 4) and

LO2 – To develop and foster awareness of the role of AEMs in the development of MT systems	
Application of AEMs	**Weaknesses of AEMs**
Gained interest in pre-trained AEMs. (P4S2)	Because AEMs have disadvantages, metrics should be combined to use their strengths. (P4S2)
Acknowledges the importance of combining sentence-level and document-level AEMs to gain more granularity in system checks. (P4S10)	Previously unaware of the weaknesses of AEMs. (P4S16)
Stated the value of a document-level AEM for being able to assign a score to discourse features. (P4S12)	
Gained interest in BERTScore. (P4S16)	

Table 7.14: Participants' views on AEM application and weaknesses within MT development (LO2), extracted from reflective reports, assignments and questionnaires.

was operationalised through lab assignments that involved the testing of AEMs on MATEO and BlonDe. Table 7.14 shows that the students were motivated to use different AEMs or acknowledged the weaknesses of the AEMs. The training aimed to focus on HE. It was never a goal to demonise AEMs, but rather, to show when and where AEMS can be applicable. In the SRL (Subsection 3.5.1.2), it was theorised that accessible AEM platforms could incentivise developers to create platforms for other evaluators to calculate their own scores, but evidence shows that accessible platforms are also useful and preferred by NLP students.

Drawing some findings from LO1, part of the reason why students were willing to combine AEMs with HE is related to students' ability to recognise the limitations of AEMs, although results from the quiz in Lesson 5 (see Subsection 6.3.5) showed that students still had trouble identifying features of reference free metrics, only correctly identifying features of reference-based metrics (Questions 2 and 3 respectively). Because less time was allocated for the specific practice of AEMs, it is understandable that students might have had more trouble operationalising or identifying features of AEMs. However, in terms of awareness, at least, there is some evidence that NLP students understand that AEMs have limitations and have uses by combining them with HE.

7.4.3 Learning Outcome 3

LO3 – To develop an understanding of key elements in the development of MT systems, including purpose, evaluation metrics, type of evaluators, and ethical considerations to ensure systems are less risky and less biased			
Evaluation Frameworks & Criteria	**UX/UI**	**Human-centred & Critical Perspectives**	**Ethical Considerations**
Affirms that guidelines are indispensable for HE. (P4S12)	Affirms that lack of consistency in UI can impact evaluations, leading to slow workflows. (P4S12)	Understands that the type of evaluator matters, and the right user/evaluator can provide meaningful feedback. (P4S5)	Expresses that biases, more specifically gender bias, should be inspected during MTQA. (P4S2, P4S7)
	Affirms that a good UI can maintain good results and prevent fatigue (P4S18)	Understands how subjectivity can be a factor to be considered in evaluation with humans. (P4S5, P4S16)	Defends that systems should benefit humans and that systems should be efficient. (P4S5, P4S10)
			Expresses concern over data governance, mentioning GDPR must be followed. (P4S12, P4S13, P4S16)
			Expresses that the TBL is essential for making ethical assessments of MT. (P4S18)

Table 7.15: Participants' views on evaluation frameworks, UX/UI, human-centred perspectives, and ethical considerations in MT development (LO3), extracted from reflective reports, assignments and questionnaires

This LO was operationalised in Units 1, 2, 3 and 4, particularly through the evaluation design assignments and reflective reports where students integrated technical, methodological, and ethical dimensions. As shown in Table 7.15, evidence from the assignments indicates that students achieved this LO by demonstrating an understanding of elements from HE and the contextual factors shaping them. The choice of evaluators, the type of feedback required and the requirements for

evaluators were discussed (P4S5 and P4S16), showing that participants understood that is not only a matter of choosing an evaluator, but what the consequences entail in choosing them. In the quiz from Lesson 5 (Subsection 6.3.5), Q7 aimed to focus on the concept of perishability, which most of the students correctly responded to. Moreover, in Lesson 6 (see subsection 6.3.6), in Task 1 students were able to identify the type of content and debate its perishability.

As discussed in Lesson 8 (see Subsection 6.3.8), students were quite vocal about their distrust towards hyperbolic claims of MT systems capabilities. That same lesson showed comments from students concerned about the transparency and reproducibility of studies. Students also developed concerns about how data is handled, as the misuse of data can lead to various risky scenarios both humans and machines. What is interesting is that some comments in Table 7.15 (P4S5 and P4S10's) mention elements very reminiscent of the TBL, which was also cited by P4S18. Although Moorkens et al. (2024a) is a position paper, perhaps it has inspired students to be aware of ethical elements in an MTQA workflow. With this evidence from the lessons and assignments, it leads me to believe that the students were successful in having an understanding of the ethical implications and at least showing concern about the ethical aspects of MT development.

7.4.4 Learning Outcome 4

Evidence for this LO emerged from assignments and observations from Units 2, 3 and 4. The thematic analysis of these assignments indicates that students were able to theorise principles from their practical tasks. When reflecting about using and evaluating MT, participants were able to notice the strengths and weaknesses of each type of metric, and thus, came to the conclusion that using both is essential. Notably, comments were made based on the idea that a comprehensive and well-rounded evaluation combines HE and AEM, and it is properly documented to be reproducible (P4S5, P4S7, P4S16). Having awareness of evaluation approaches and different evaluators, some students considered the suitability of these factors in the

LO4 – To design replicable evaluation of MT systems, cognisant of diverse evaluation approaches and types of evaluators		
Reflections on MT use		
Reproducibility & Comprehensive Evaluation	**Domain and Use Cases**	**Ethics and Evaluation Design**
Knowledge of metrics and MTQA allows the appropriate use of evaluators and metrics for reproducible experiments. (P4S5)	Understanding of how MT works is essential to adapt for different use cases (P4S10)	A well-detailed evaluation plan and the use of ethics is the right combination to understand how an MT system works. (P4S18)
Knowledge of both types of metrics allows the comprehensive evaluation of results of MT. (P4S7)	Domain appropriateness and professional expertise are essential to understand and interact with MT. (P4S17)	
Knowledge of both HE and AEMs with focus on reproducibility is necessary for a comprehensive understanding of MT. (P4S16)		

Table 7.16: Participants' reflections on reproducibility, domain use, and ethical evaluation design in MT development (LO4), extracted from reflective reports, assignments and questionnaires.

design of MTQA. In particular, P4S10 and P4S17 believe that these aspects are essential and must be documented. Likewise, P4S18 believes that it is ethical to have a well-detailed evaluation plan.

While Lesson 6 (see subsection 6.3.6) provided an exercise in documenting most aspects of MTQA, it did not emphasise much replicability. Lesson 7 (see subsection 6.3.7 informed students of the ethical aspects and replicability was one of the topics, which could have provided some awareness of the dangers of lacking well-documented processes. Therefore, it is appropriate to state that there is some minor evidence that elements of LO4 were achieved, but I do not believe that it was fully achieved.

7.4.5 Learning Outcome 5

Evidence from this LO emerged from assignments and observations from Unit 4. In Lesson 8 (see Section 6.3.8) it was possible to observe that the students were

LO5 – To report results from the evaluation of an MT system addressing the context, type of evaluators, and use case	
Caution with MT use	
Reliability	**Consequences of poor MTQA**
Show concern about the reliability of MT systems. (P4S5, P4S7, P4S10, P4S11)	Show concern over poor MTQA because of risks, such as misinformation and biases. (P4S2, P4S13)
Shows concern about the validation of results of MT systems with poor evaluations. (P4S18)	MT development can be slowed down if evaluation methods are not followed properly. (P4S16)

Table 7.17: Participants' views on reliability and the consequences of poor MTQA within MT evaluation reporting (LO5), extracted from reflective reports, assignments and questionnaires.

growing concerned about claims of MT systems, which arise from Task 1. It was more evident when analysing the posttest questionnaires, where students showed reservations about the reliability of MT systems (P4S5, P4S7, P4S10, P4S11) and the validation of results of poor MT evaluations (P4S18). Students are concerned about the risks of MT through poor quality assurance, mainly misinformation and propagation of biases (P4S2, P4S13). One of those risks is also the idea that the development of MT can be slowed if MTQA is not performed properly, according to a student (P4S16). At the end of the course, the students were able to not only identify what they perceived as flaws, but also to explain their reasoning. Furthermore, students showed active concern about the consequences of flawed MT reporting, leading to the belief that they are at least aware of the consequences of bad reporting of results. In general, the findings of the course suggest that the training effectively supported the development of LO1 through LO5, particularly through critical reflection and ethical awareness.

Chapter 8

Conclusion

In this chapter, I present the conclusions of my research, addressing each RQ presented in chapter 1. The general contributions of this thesis are summarised in 8.2. The limitations of this research are discussed in 8.3; and future work in 8.4.

8.1 Research Questions Revisited

8.1.1 RQ1 - What can be taught to NLP students to encourage them to carry out HE?

This PhD thesis demonstrated that the use of ADDIE, combined with quantitative and qualitative data collection methods in the classroom, resulted in training that produced evidence that students are willing to adopt HE and are more confident than before in using HE in their own practice, with statistically significant results. Although these are the results to a course applied for a specific cohort in specific educational conditions (factoring duration of the course, time of course delivery, the pedagogical design), some materials and exercises show potential to be reproduced in other contexts. It has been shown that exercises that put developers-in-training in the position of translators and other end-users as evaluator can promote changes in students' attitudes about MT and its evaluation. Throughout the training course delivered, students consistently reported the usefulness of tools and how they had

been able to apply the knowledge gained in core modules of their master's programmes. However, what also made students willing to adopt HE in their practice was the teaching of AEMs, their use and their weaknesses. Although not the focus of this investigation, students' confidence with AEMs also increased, with statistically significant results as compiled in Appendix I. Students reported that they were previously unaware of the shortcomings of the AEMs and understanding these shortcomings highlighted the areas where HE plays an important role in MTQA.

Good methodological practices that are already adopted in the research landscape appear to provide a solid basis for students to make decisions, potentially easing the difficulties of identifying and designing an effective workflow. Students reported that classroom exercises that tested the design of MTQA are complex but have proven to be important in their learning process. Teaching ethics and sustainability also helped shape students' attitudes towards the consequences of "good" or "bad" MTQA practices. Although these practices may have led to distrust of MT in terms of unethical MTQA development or MT systems that can potentially cause harm, I believe this fear is awareness of what can happen when HE is not carried out in scenarios in which it should be implemented.

In general, the thesis presents content and areas that show promise in helping students demonstrate both confidence and positive attitudes that imply the intention to use HE. Several participating students reported that the teaching practices contributed to their learning, but they also mentioned weaknesses in teaching with comments stating the difficulty of more technical aspects. Those difficulties will be further discussed in the section 8.3 related to limitations.

8.1.2 RQ2 - What content related to MTQA is currently taught in post-graduate NLP and TS programmes, if any?

This RQ is pursued through RQ2a and RQ2b.

8.1.2.1 RQ2a - What research has been published on MTQA teaching for TS and NLP?

This RQ was answered through the SRL performed in Chapter 3. A substantial number of studies from TS were identified and coded as teaching practices, curricular & syllabi insights, resources, insights on modality and recommendations from the literature that serve as rationale for good MTQA teaching. Through the codification of these studies into themes, it was possible to define principles that were used to design and develop the course proposed in this thesis.

When identifying MTQA teaching through the lens of TS, it was found that the technical and linguistic aspects of MT had to be understood so that translators could use MT to the fullest. In parallel, for NLP education this means showing the complex dimension of language to computer scientists. In other words, world knowledge, context and cultural aspects are facets that are relevant for MT systems and that MT might struggle to replicate. That is what exercises from the perspective of translators and end-users intended to show to computer scientists.

Other factors found from TS had shown that no matter the modality, MTQA training should be flexible and should be adapted to the context, including the type of students, type of training offered. The rationales from TS highlight that not only are the technical aspects of MT important, but the ethical elements and interdisciplinary approaches also show benefit to translators. For effective training, in the context of this thesis, interdisciplinary approaches are effective, providing more reasons to implement TS tools & approaches into NLP education. Finally, the literature in TS education presents effective examples of pedagogical resources for translators; those resources can serve as a great source of inspiration for developers to design intuitive evaluation platforms.

8.1.2.2 RQ2b - What has been taught in MTQA for NLP postgraduate students?

The curricular analysis performed for this RQ (section 3.7) was meant to complement RQ2a. It revealed a prevalence of explicit MTQA evaluation when MT was present in the curriculum. More than that, it showed that it is difficult to dedicate modules exclusively to evaluation across NLP tasks, not just MT. Modules on evaluation are sometimes optional, including modules that focus on MT as a task. The curricular analysis highlighted a potential scarcity of structured educational content on MTQA education for NLP postgraduate students, highlighting the importance of the type of study conducted in this thesis.

8.1.3 RQ3 - What can be drawn from TS when teaching MTQA for NLP postgraduate audiences?

This RQ was partially answered in Phase 1c (section 4.1), identifying the content and recommendations of TS and NLP educators, as the aim was to ask educators about topics that would be indispensable for the NLP students studying MTQA. Then, in the evaluation of the course (section 7.3), I found evidence that the students wanted to use tools and methods popular in TS or that originated from TS. Data have shown that certain HE metrics, such as adequacy and fluency, were consistently adopted, in addition to reliability measures such as the use of IAA and document-level evaluation. What also appeared to be popular was the use of Jupyter notebooks and the use of MATEO, as students praised platforms that streamlined the study or calculation of AEM scores. Finally, it is noteworthy to mention that ethical aspects regarding, for example, data management can be perceived in students' concerns and sustainability can also be noticed in their attitudes with the introduction of the TBL framework.

8.1.4 RQ4 - How effective is the content proposed for HE in an MT evaluation module?

To respond to this RQ, the LOs proposed in the course had to be verified. LO1 aimed to equip students with the awareness of HE's role in MTQA, which was shown with evidence of how HE is applied, how it interacts with AEMs and what frameworks and criteria can be adopted to reduce ambiguity and improve UX for the evaluators. LO2 aimed to equip students with the awareness of AEMs' role in MTQA, which was shown with evidence of how AEMs are applied and their weaknesses. LO3 aimed to equip students with the understanding of key elements in MTQA, which was shown with evidence related to evaluation frameworks and criteria, UX/UI, perspectives that are human-centred and/or critical and ethical considerations. LO4 aimed to equip students with the ability to design replicable evaluation of MT, which was shown with evidence related to the topic of reproducibility, comprehensiveness in evaluation, use cases and ethics. LO5 aimed to equip students with the ability to report the results of MTQA, which was shown with evidence of students focusing on the reliability of MTQA and what the consequences are of poor MTQA and its reporting. This research has shown that students were able to demonstrate through their reflections and assignments how they would handle aspects that pertain to each LO appropriately.

8.2 Contributions

This PhD thesis aimed to contribute to MTQA pedagogy, particularly within NLP education, by responding to a pressing challenge in contemporary NLP and MT research. MT systems and AI-driven technologies, while increasingly powerful and widely deployed, remain prone to producing and amplifying biases, errors, and misleading outputs (Koehn and Knowles, 2017; Guerreiro et al., 2023; Bowman, 2024; Huang et al., 2025). As discussed in Chapter 1, the rapid adoption of these systems has been accompanied by recurring hype cycles in which capabilities of those

systems are overstated and weaknesses are either masked or insufficiently analysed (Moorkens, 2022). This creates conditions in which evaluation practices that are narrow, poorly designed, or overly reliant on AEMs risk giving a distorted view of MT systems' performance. Such pitfalls are well documented in the literature and become especially consequential in high-stakes domains, where an insufficient, misguided, or improper evaluation can have significant social, economical, ethical, or legal implications (Marie et al., 2021; Vieira et al., 2020).

Regulation of emerging technologies can be difficult. However, a proactive and scalable response lies in training the next generation of NLP professionals to perform responsible, comprehensive, and above all critical evaluation. By equipping students with conceptual and methodological tools to examine both HE practices and AEMS, teaching can be a key factor in minimising some of the existing hype and misguided claims of MT systems, ultimately supporting more accountable MT and AI development in general. This thesis filled this gap with the following key contributions.

1. Structured diagnosis and identification of educational resources in MTQA through Phases 1a, 1b, 1c.

2. Structured development and planning of pedagogical materials and resources of a training course on critical MTQA through Phases 2 and 3.

3. Structured reporting of the course delivery and evaluation of the effectiveness of a critical training course on MTQA through Phase 4 and 5.

The first contribution produced an account in Chapter 3 of what is currently taught, which resources exist, and where the most significant pedagogical gaps are located. The findings show that TS already contains established practices relevant to MTQA, while NLP does not employ equivalent structures in its pedagogy. This contribution identifies the gap clearly and shows how TS resources can support the design of MTQA training for NLP learners. The accompanying survey (Section 4.1) strengthens this contribution by identifying content that educators consider suitable

for inclusion in an MTQA course and by confirming the direction that informed the subsequent phases of this thesis.

The second contribution consists of the creation of pedagogical materials and methodologies that respond directly to the need for responsible and critical MTQA. These materials draw from TS and are structured through the ADDIE framework, which facilitated clear separation between phases of design and development. The materials can be used or adapted at different levels of granularity, including units, lessons, or individual tasks. The use of Freirean critical pedagogy positions the teacher to promote a reflective stance towards MT systems and their evaluation, encouraging students to examine assumptions, procedures, and consequences of evaluation choices.

The third contribution is achieved through detailed documentation of the course in practice and through evaluation of its pedagogical effectiveness. This includes observation of how students engage with the materials, identification of challenges that arise during teaching, and evidence of how students begin to question and extend MTQA practices. The results indicate increased confidence in the use of both HE and AEMs and show that the course is effective for the NLP audiences. Because students feel confident and compelled to adopt the resources they acquired during the training, this shows that MTQA is useful for NLP students and can be used for other tasks.

These contributions show that responsible and critical MTQA training is achievable in practice and can be implemented in a structured and pedagogically coherent way. This is particularly important at a time when MT remains far from a solved task and when certain languages and language pairs continue to lack sufficient resources (Moorkens et al., 2024b). At the same time, regulations are gradually emerging that describe clear requirements for developers and providers, including the need to document evaluation methodologies, identify system limitations, and ensure transparency and safety procedures as shown in Annex IX of the AI Act (European Parliament and the Council of the European Union, 2024). These devel-

opments reinforce the need to prepare students to understand and act in accordance with such expectations. The training course developed in this thesis promotes this form of preparation.

For these reasons, the course has shown the potential to extend beyond MT. The structure of the course, the critical emphasis, and the integration of best practices from the literature form a solid foundation for preparing NLP students to evaluate AI systems more broadly. As evaluation becomes an increasingly important component of regulatory compliance and responsible development, this thesis offers a timely and relevant contribution that supports the formation of critically informed NLP professionals capable of developing and assessing AI systems responsibly.

8.3 Limitations

Although this PhD thesis presented a novel course on MTQA with meticulous planning based on the ADDIE framework, some challenges and limitations were found in different phases of this research. In Phase 1b, the curricular analysis had two flaws. The first, related to scope is that it represented only European postgraduate NLP programmes. Although it is pertinent to be realistic with deliverables, it is inevitable that some perspectives from other geographical regions where MTQA teaching may be present are left undocumented. Second, the analysis relied exlcusively on publicly available data. While the data found were rich, different universities offered information about their programmes to varying degrees, making it difficult to acquire data in a standard manner. Some information was only acquired by visiting very specific and unintuitive sections of university websites, while other pieces of information were not able to be extracted due to the university limiting the amount of available information for various reasons.

Phase 2 relied on resources related to my institution and personal affiliations (such as the ADAPT centre), which may not be transferable to the implementation of this course into other contexts. It is expected that any training course requires adaptation to be taught in another context, and the literature even supports this,

but there might be hurdles in the teaching of the course if, for example, the classroom design and resources differ. Furthermore, it based on an idealised profile of students, which may have stemmed from my own assumptions, for example that NLP students would have an inherent knowledge of AEMs.

Phase 4 revealed that some choices in the course design, shown in Phase 3, may have introduced challenges in the students' learning process. Most notably, the technical difficulty with AEMs in Unit 2, Lesson 4 and lack of time for more focus on AEMs and IAA measures other than Cohen's Kappa may have revealed that these topics could require dedicated labs to be explored fully by students. Other elements certainly required more time. In Unit 3, Lesson 5 and 6, for example, statistical measures to correlate results of HE and AEM scores could have been addressed, but were present only as part of an instruction file for students. In addition, the lack of student data in Unit 3 may have demonstrated that the learning activities lacked components in which students could freely comment on aspects of evaluation design. By itself, it is not flawed to not contain such exercises, but it makes it more difficult to diagnose and analyse students' progress and challenges based on classroom observation and field notes alone. These pedagogical obstacles perhaps could have been remedied had the syllabus of the course and the course itself been presented to educators from Phase 1c, to validate the lesson planning, materials and learning activities. Furthermore, as the course was created over the past four years, it was focused mostly on NMT and its output. An attempt was made to include the output of LLMs as an option for students where relevant, but the course was clearly not focused on LLM-driven translation.

Phase 5, focused on evaluation, showed promising evidence that suggests that students can employ HE and are willing to use it, but considering the short amount of time, its conclusions are still limited. Ideally, a longitudinal study with a longer training course or with follow-up questionnaires or interviews could further investigate how students perceive HE, how confident they have felt and if they had implemented MTQA approaches and tools in MT development or any other NLP task.

In addition, another angle not explored was a course with modifications based on feedback from students to be presented to Phase 1c educators, as a way to validate the training course after its first iteration.

Another limitation present due to the scope of this PhD thesis is the lack of a full analysis for TT students who took part in this training. Considering their difficulties and how the course was not adapted to their needs, I believe it would not be a representative analysis of the full potential of a translator-in-training. As such, it would not be appropriate to check the confidence of these students or to understand how effective the training was.

8.4 Future Work

As this was a successful attempt to propose an MTQA course for NLP audiences, feedback from students suggests expanding the course in content and scope. It would be valuable to include and test MTQA methods with LLM-driven translation or other tasks, to determine changes in workflows, methodologies and experiment with additional constraints, such as prompts. Furthermore, the training course explored domains from a generalised perspective. It is possible to tailor future editions of the course or units to specialised domains, such as the legal domain or medical domain. Implementing MTQA in specialised domains presents more or less risks depending on the type of content, as evidenced by the concept of perishability. It would be worthwhile to investigate to what extent MTQA approaches, tools, and methods should be taught in a different manner or not when LLMs are considered.

Other aspects of MTQA appeared to be of interest to students, as evidenced by their comments and reflections. With the addition of UX elements that affect the evaluation process, it is a promising field of study to determine to what extent and how UX should be taught to computer scientists, something that Briva-Iglesias et al.'s (2023) research suggested as another factor that affects MT development. Another factor that must gain more attention is the teaching of sustainability, primarily with the use of the TBL. The TBL itself is not a pedagogical framework

nor was it specifically envisioned for classroom use, but students' comments and feedback suggest that they recognised the value of formalising the concept of sustainability through the framework. More pedagogical applications of the TBL must be implemented in order to verify whether it is possible to make MT development more sustainable by including learning activities inspired or informed by the TBL.

Another promising venue of investigation is adapting the course in terms of tech-friendliness for TS audiences, specifically students from TT. While they were present in this cohort, it is most appropriate to tailor certain learning activities and explanations, such as AEMs and the use of Python for the calculation of IAA in order for translators to serve as important stakeholders in MTQA. They already receive training to a great extent in linguistic and technical aspects, as the literature demonstrates. However, because this is a full module focused on evaluation, with emphasis on the perspective of developers, this could provide translators with more resources and the right language to communicate and advocate for the appropriate use of HE informed by translation expertise.

Bibliography

ACL Anthology (2023), 'Workshop on Teaching Natural Language Processing'. Accessed on: 11 September 2023.
URL: *https://aclanthology.org/venues/teachingnlp/*

Allen, B., McGough, A. S. and Devlin, M. (2021), 'Toward a Framework for Teaching Artificial Intelligence to a Higher Education Audience', *ACM Transactions on Computing Education (TOCE)* **22**(2), 1–29.

Alm, C. O., Womack, K., Haake, A. and Engström, T. (2016), 'A Pedagogical Model for Computational Linguistics across Curricular Boundaries', *Language and Linguistics Compass* **10**(7), 335–345.

Amrhein, C. and Sennrich, R. (2022), Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET, *in* Y. He, H. Ji, S. Li, Y. Liu and C.-H. Chang, eds, 'Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)', Association for Computational Linguistics, Online only, pp. 1125–1141.
URL: *https://aclanthology.org/2022.aacl-main.83/*

Andreata, M. A. (2019), 'Aula Expositiva e Paulo Freire', *Ensino em Re-Vista* **26**(3), 700–724.
URL: *https://seer.ufu.br/index.php/emrevista/article/view/50981*

Artemova, E., Apishev, M., Kirianov, D., Sarkisyan, V., Aksenov, S. and Serikov,

O. (2021), Teaching a massive open online course on natural language processing, *in* 'Proceedings of the Fifth Workshop on Teaching NLP', Association for Computational Linguistics, Online, pp. 13–27.
URL: *https://aclanthology.org/2021.teachingnlp-1.2/*

Ary, D., Jacobs, L. C., Sorensen, C. and Razavieh, A. (2010), *Introduction to Research in Education*, 8th edn, Wadsworth Cengage Learning, Belmont, CA.

Asscher, O. (2023), 'The position of machine translation in translation studies: A definitional perspective', *Translation Spaces* **12**(1), 1–20.

ASTM International (2023), *Standard Practice for Quality Assurance in Translation.*
URL: *https://cdn.standards.iteh.ai/samples/114457/435bccaf67ee4c37ada19c603fdd0395/ASTM-F2575-23.pdf*

Banerjee, S. and Lavie, A. (2005), METEOR: An automatic metric for mt evaluation with improved correlation with human judgments, *in* 'Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization', Association for Computational Linguistics, Ann Arbor, Michigan, USA, pp. 65–72.
URL: *https://aclanthology.org/W05-0909/*

Barg, M., Fekete, A., Greening, T., Hollands, O., Kay, J. and Kingston, J. H. (2000), 'Problem-Based Learning for Foundation Computer Science Courses', *Computer Science Education* **10**(2), 99–106.
URL: *https://doi.org/10.1076/0899-3408(200008)10:2;1-C;FT109*

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M. and Zampieri, M. (2019), Findings of the 2019 conference on machine translation (WMT19), *in* O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, M. Turchi and K. Verspoor, eds,

'Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)', Association for Computational Linguistics, Florence, Italy, pp. 1–61.
URL: *https://aclanthology.org/W19-5301/*

Bart, A. C., Kafura, D., Shaffer, C. A. and Tilevich, E. (2018), Reconciling the Promise and Pragmatics of Enhancing Computing Pedagogy with Data Science, *in* 'Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE '18)', ACM, pp. 1028–1033.
URL: *https://doi.org/10.1145/3159450.3159465*

Bar-Hillel, Y. (1959), Report on the State of Machine Translation in the United States and Great Britain, Technical Report Technical Report No. 1, Hebrew University, Jerusalem, Israel. Prepared for the United States Office of Naval Research, Information Systems Branch, under Contract No. NONR-2578(00), N.R.049-130.
URL: *https://www.mt-archive.net/50/Bar-Hillel-1959.pdf*

Bawden, R., Sennrich, R., Birch, A. and Haddow, B. (2018), Evaluating Discourse Phenomena in Neural Machine Translation, *in* 'Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)', Association for Computational Linguistics, New Orleans, Louisiana, pp. 1304–1313.
URL: *https://aclanthology.org/N18-1118*

Berssanette, J. H. and de Francisco, A. C. (2021), 'Active Learning in the Context of the Teaching/Learning of Computer Programming: A systematic review', *Journal of Information Technology Education: Research* **20**, 201–220.
URL: *https://www.informingscience.org/Publications/4767*

Bhattacharyya, P., Chatterjee, R., Freitag, M., Kanojia, D., Negri, M. and Turchi, M. (2023), Findings of the WMT 2023 shared task on automatic post-editing, *in* P. Koehn, B. Haddow, T. Kocmi and C. Monz, eds, 'Proceedings of the Eighth

Conference on Machine Translation', Association for Computational Linguistics, Singapore, pp. 672–681.
URL: *https://aclanthology.org/2023.wmt-1.55/*

Blumczyński, P. (2016), *Ubiquitous Translation*, Routledge Advances in Translation and Interpreting Studies, Routledge, London & New York.

Bogdan, R. C. and Biklen, S. K. (2006), *Qualitative Research for Education: An Introduction to Theories and Methods*, 5th edn, Pearson.

Bogush, A. M., Korolova, T. M. and Popova, O. V. (2019), 'Teaching machine translation to the students majoring in the humanities', *Information Technologies and Learning Tools* **71**(3), 122–136.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L. and Turchi, M. (2017), Findings of the 2017 conference on machine translation (WMT17), *in* O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn and J. Kreutzer, eds, 'Proceedings of the Second Conference on Machine Translation', Association for Computational Linguistics, Copenhagen, Denmark, pp. 169–214.
URL: *https://aclanthology.org/W17-4717/*

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K. and Zampieri, M. (2016), Findings of the 2016 conference on machine translation, *in* O. Bojar, C. Buck, R. Chatterjee, C. Federmann, L. Guillou, B. Haddow, M. Huck, A. Jimeno Yepes, A. Névéol, M. Neves, P. Pecina, M. Popel, P. Koehn, C. Monz, M. Negri, M. Post, L. Specia, K. Verspoor, J. Tiedemann and M. Turchi, eds, 'Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers', Association for Computational Linguistics,

Berlin, Germany, pp. 131–198.
URL: *https://aclanthology.org/W16-2301/*

Bortoni-Ricardo, S. M. (2008), *O Professor Pesquisador: Introdução à Pesquisa Qualitativa*, Parábola Editorial, São Paulo, Brasil.

Bowker, L. and Ciro, J. B. (2019), *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*, Emerald Publishing Limited.

Bowman, S. R. (2024), 'Eight Things to Know About Large Language Models', *Critical AI* **2**(2).
URL: *https://read.dukeupress.edu/critical-ai/article/doi/10.1215/2834703X-11556011/400182/Eight-Things-to-Know-about-Large-Language-Models*

Brady, L. (2013), 'Curriculum Evaluation: Where are we now?', *Curriculum and Teaching* **28**(2), 61–72.

Brew, C., Dickinson, M. and Meurers, D. (2005), "Language and Computers": Creating an introduction for a general undergraduate audience, *in* 'Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL', Association for Computational Linguistics, pp. 15–22.
URL: *https://aclanthology.org/W05-0103/*

Breyel, C. and Grass, T. (2020), 'Machine translation and" biotranslation": an unlikely marriage.', *SKASE Journal of Theoretical Linguistics* **17**(4).

Briva-Iglesias, V., Camargo, J. L. and Dogru, G. (2024), 'Large Language Models "Ad Referendum": How Good Are They at Machine Translation in the Legal Domain?', *MonTI. Monografías de Traducción e Interpretación* **16**(2), 45–60.
URL: *https://www.e-revistes.uji.es/index.php/monti/article/view/7514*

Briva-Iglesias, V., O'Brien, S. and Cowan, B. R. (2023), 'The impact of traditional and interactive post-editing on machine translation user experience, quality, and productivity', *Translation, Cognition & Behavior* **6**(1), 60–87.

Brown, A. H. and Green, T. D. (2024), *The Essentials of Instructional Design: Connecting Fundamental Principles with Process and Practice*, Taylor & Francis, Abingdon, UK.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020), Language Models Are Few-Shot Learners, *in* 'Advances in Neural Information Processing Systems', Vol. 33, pp. 1877–1901.
URL: *https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html*

Bulut, S. Ö. (2019), 'Integrating machine translation into translator training: towards 'human translator competence'?', *transLogos Translation Studies Journal* **2**(2), 1–26.

Canfora, C. and Ottmann, A. (2018), 'Of Ostriches, Pyramids, and Swiss Cheese: Risks in Safety-Critical Translations', *Translation Spaces* **7**(2), 167–201.

Canfora, C. and Ottmann, A. (2020), 'Risks in Neural Machine Translation', *Translation Spaces* **9**(1), 58–77.
URL: *https://www.jbe-platform.com/content/journals/10.1075/ts.00021.can*

Cao, C., Ding, Z., Lee, G.-G., Jiao, J., Lin, J. and Zhai, X. (2023), Elucidating STEM Concepts through Generative AI: A Multi-modal Exploration of Analogical Reasoning, *in* 'Proceedings of the Sixth Workshop on Teaching NLP'.
URL: *https://arxiv.org/abs/2308.10454*

Castilho, S. (2020), On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation, *in* 'Proceedings of the Fifth Conference on Machine Translation', Association for Computational Linguistics, Online, pp. 1150–1159.
URL: *https://aclanthology.org/2020.wmt-1.137*

Castilho, S. (2021), Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation, *in* 'Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)', Association for Computational Linguistics, Online, pp. 34–45.
URL: *https://aclanthology.org/2021.humeval-1.4*

Castilho, S. and Caseli, H. d. M. (2024), Tradução Automática: Abordagens e Avaliação, *in* H. M. Caseli and M. G. V. Nunes, eds, 'Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português', 2 edn, BPLN, book chapter 21.
URL: *https://brasileiraspln.com/livro-pln/2a-edicao/parte-aplicacoes/cap-mt/cap-mt.html*

Castilho, S., Cavalheiro Camargo, J. L., Menezes, M. and Way, A. (2021), DELA corpus - a document-level corpus annotated with context-related issues, *in* L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-Jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzmán, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita and C. Monz, eds, 'Proceedings of the Sixth Conference on Machine Translation', Association for Computational Linguistics, Online, pp. 566–577.
URL: *https://aclanthology.org/2021.wmt-1.63/*

Castilho, S., Doherty, S., Gaspari, F. and Moorkens, J. (2018), Approaches to human and machine translation quality assessment, *in* J. Moorkens, S. Castilho, F. Gaspari and S. Doherty, eds, 'Translation Quality Assessment: From Principles to Practice', Springer International Publishing, Cham, Switzerland, pp. 9–38.
URL: *https://doi.org/10.1007/978-3-319-91241-7_2*

Castilho, S. and Knowles, R. (2024), 'A Survey of Context in Neural Machine Translation and Its Evaluation', *Natural Language Processing* **31**(4), 986–1016.
URL: *https://www.cambridge.org/core/journals/natural-language-*

processing/article/survey-of-context-in-neural-machine-translation-and-its-evaluation/875C3E2DEAAB8845A0F05988D5A18B55

Castilho, S., Popović, M. and Way, A. (2020), On Context Span Needed for Machine Translation Evaluation, *in* N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis, eds, 'Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)', European Language Resources Association (ELRA), Marseille, France, pp. 3735–3742.
URL: *https://aclanthology.org/2020.lrec-1.460*

Catford, J. C. (1965), *A Linguistic Theory of Translation: An Essay in Applied Linguistics*, Oxford University Press, London.

Cid-Leal, P., Espín-García, M.-C. and Presas, M. (2019), 'Traducción automática y posedición: Perfiles y competencias en los programas de formación de traductores', *MonTI: Monografías de Traducción e Interpretación* **11**(0), 187–214.

Cochran-Smith, M. and Lytle, S. L. (2009), *Inquiry as Stance: Practitioner Research for the Next Generation*, Teachers College Press, New York, NY.
URL: *https://www.tcpress.com/inquiry-as-stance-9780807749708*

Cohen, J. (1960), 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement* **20**(1), 37–46.
URL: *https://journals.sagepub.com/doi/10.1177/001316446002000104*

Cohen, L., Manion, L. and Morrison, K. (2018), *Research Methods in Education*, 8th edn, Routledge, New York, NY.

Cohen, P. R. (1995), *Empirical Methods for Artificial Intelligence*, MIT Press, Cambridge, MA.
URL: *https://mitpress.mit.edu/9780262534178/empirical-methods-for-artificial-intelligence/*

Connelly, F. M. and Xu, S. J. (2010), An Overview of Research in Curriculum Inquiry, *in* P. Peterson, E. Baker and B. McGaw, eds, 'International Encyclopedia of Education', Elsevier, Oxford, UK.

Córdova-Esparza, D.-M., Romero-González, J.-A., Córdova-Esparza, K.-E., Terven, J. and López-Martínez, R.-E. (2024), 'Active Learning Strategies in Computer Science Education: A systematic review', *Multimodal Technologies and Interaction* **8**(6), 50.
URL: *https://www.mdpi.com/2414-4088/8/6/50*

Costa-jussà, M., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J. et al. (2022), 'No language left behind: Scaling human-centered machine translation', *arXiv preprint arXiv:2207.04672* .
URL: *https://arxiv.org/abs/2207.04672*

Creswell, J. W. (2015), *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*, 5th edn, Pearson, Boston, MA.

Creswell, J. W. and Clark, V. L. P. (2017), *Designing and Conducting Mixed Methods Research*, 3rd edn, Sage Publications, Thousand Oaks, CA.

Creswell, J. W. and Creswell, J. D. (2023), *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 6th edn, SAGE Publications, Thousand Oaks, CA.

Cronin, M. (2019), Translation, Technology and Climate Change, *in* M. O'Hagan, ed., 'The Routledge Handbook of Translation and Technology', Routledge, London & New York, pp. 516–530.

Cui, R. and Teo, P. (2020), 'Dialogic Education for Classroom Teaching: A critical review', *Language and Education* **34**(1), 1–17.

Dejica-Cartis, D. (2012), 'Developing the electronic tools for translators syllabus

at politehnica university of timisoara', *Procedia-Social and Behavioral Sciences* **46**, 3614–3618.

Diamond, R. M. (2008), *Designing and Assessing Courses and Curricula: A Practical Guide*, John Wiley & Sons, San Francisco, CA.

Dick, W., Carey, L. and Carey, J. O. (2021), *The Systematic Design of Instruction*, 9th edn, Pearson, Hoboken, NJ.

Dignum, V. (2020), Responsibility and artificial intelligence, *in* 'The Oxford Handbook of Ethics of AI', Oxford University Press, Oxford, UK, pp. 214–231.

Dobreva, R., Zhou, J. and Bawden, R. (2020), Document Sub-structure in Neural Machine Translation, *in* 'Proceedings of the Twelfth Language Resources and Evaluation Conference', European Language Resources Association, Marseille, France, pp. 3657–3667.
URL: *https://aclanthology.org/2020.lrec-1.451*

Doherty, S. (2017), Issues in Human and Automatic Translation Quality Assessment, *in* D. Kenny, ed., 'Human Issues in Translation Technology', Routledge, London & New York, pp. 131–148.

Doherty, S. and Kenny, D. (2014), 'The design and evaluation of a statistical machine translation syllabus for translation students', *The Interpreter and Translator Trainer* **8**(2), 295–315.

Doherty, S., Moorkens, J., Gaspari, F. and Castilho, S. (2018), On education and training in translation quality assessment, *in* J. Moorkens, S. Castilho, F. Gaspari and S. Doherty, eds, 'Translation Quality Assessment: From Principles to Practice', Springer International Publishing, Cham, Switzerland, pp. 95–106.
URL: *https://doi.org/10.1007/978-3-319-91241-7_5*

Doolittle, P., Wojdak, K. and Walters, A. (2023), 'Defining Active Learning: A restricted systematic review', *Teaching & Learning Inquiry* **11**, 1–17.
URL: *https://journalhosting.ucalgary.ca/index.php/TLI/article/view/76484*

Dousay, T. A. and Logan, R. (2012), Analyzing and Evaluating the Phases of ADDIE, *in* 'Proceedings of the 2012 Annual Meeting of the Georgia Educational Research Association'.
URL: *https://www.learninginterest.com/ADDIE.pdf*

Duschl, R., Maeng, S. and Sezen, A. (2011), 'Learning Progressions and Teaching Sequences: A review and analysis', *Studies in Science Education* **47**(2), 123–182.
URL: *https://www.tandfonline.com/doi/abs/10.1080/03057267.2011.604476*

Elkington, J. (1997), *Cannibals with Forks: The Triple Bottom Line of 21st Century Business*, Capstone Publishing, Oxford, UK.

Esqueda, M. D. (2021), 'Machine Translation: Teaching and Learning Issues', *Trabalhos em Linguística Aplicada* **60**(1), 282–299.

European Parliament and the Council of the European Union (2024), *Regulation (EU) 2024/1689 – Artificial Intelligence Act*, Luxembourg.
URL: *https://eur-lex.europa.eu/eli/reg/2024/1689/oj*

Farrell, M. et al. (2017), Building a custom machine translation engine as part of a postgraduate university course: a case study, *in* 'Proceedings of the 39th Conference Translating and the Computer', pp. 35–39.

Federmann, C. (2018), Appraise: An open-source toolkit for manual evaluation of mt output, *in* 'Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations', Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 86–88.
URL: *https://aclanthology.org/C18-2019/*

Fei, L., Yunus, M. M., Tang, H. and Fan, C. (2024), 'A Systematic Review of STEM Teaching in Higher Education (2014–2023)', *International Journal of Academic Research in Progressive Education and Development* **13**(3), 3217–3231.

Fielding, N. G., Blank, G. and Lee, R. M., eds (2016), *The SAGE Handbook of Online Research Methods*, SAGE Publications Ltd, London, UK.

Filatro, A. (2023), *Design Instrucional para Professores*, Editora Senac São Paulo, São Paulo, Brasil.

Filatro, A. C. and Bileski, S. M. C. (2017), *Produção de Conteúdos Educacionais*, Saraiva Educação SA, São Paulo, Brasil.

Fleiss, J. L. (1971), 'Measuring nominal scale agreement among many raters', *Psychological Bulletin* **76**(5), 378–382.
URL: *https://doi.org/10.1037/h0031619*

Florez, S. and Alcina, A. (2011), 'Free/open-source software for the translation classroom: A catalogue of available tools', *The Interpreter and Translator Trainer* **5**(2), 325–357.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. and Vayena, E. (2018), 'AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations', *Minds and Machines* **28**(4), 689–707.

Freeman, D. (1998), *Doing Teacher Research: From Inquiry to Understanding*, Heinle & Heinle, Boston, MA.
URL: *https://eltngl-nextjs-proxy.prod.eks.cengage.io/products/9780838479001*

Freire, P. (1996), *Pedagogia da Autonomia: Saberes Necessários à Prática Educativa*, Coleção Leitura, 25 edn, Paz e Terra, São Paulo, Brasil.

Gilbert, D. (2023), Recent claims of human–machine parity in translation highlight core issues surrounding the human evaluation of machine translation, *in* I. Lacruz, ed., 'Translation in Transition: Human and Machine Intelligence', John Benjamins Publishing Company, pp. 83–99.
URL: *https://www.degruyter.com/document/doi/10.1075/ata.xx.05gil/html*

Gong, Z., Zhang, M. and Zhou, G. (2015), Document-level machine translation evaluation with gist consistency and text cohesion, *in* B. Webber, M. Carpuat,

A. Popescu-Belis and C. Hardmeier, eds, 'Proceedings of the Second Workshop on Discourse in Machine Translation', Association for Computational Linguistics, Lisbon, Portugal, pp. 33–40.
URL: *https://aclanthology.org/W15-2504/*

González Bermúdez, M., Giménez, J. and Màrquez Villodre, L. (2012), A graphical interface for mt evaluation and error analysis, *in* 'Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, pp. 139–144.

González Pastor, D. (2021), 'Introducing Machine Translation in the Translation Classroom: A Survey on Students' Attitudes and Perceptions', *Revista Tradumàtica: Tecnologies de la Traducció* (19), 47–65.

Gough, D., Oliver, S. and Thomas, J. (2017), *An Introduction to Systematic Reviews*, 2nd edn, SAGE Publications Ltd, London, UK.

Graham, Y., Baldwin, T., Moffat, A. and Zobel, J. (2013), Continuous Measurement Scales in Human Evaluation of Machine Translation, *in* A. Pareja-Lora, M. Liakata and S. Dipper, eds, 'Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse', Association for Computational Linguistics, Sofia, Bulgaria, pp. 33–41.
URL: *https://aclanthology.org/W13-2305/*

Graham, Y., Haddow, B. and Koehn, P. (2019), 'Translationese in machine translation evaluation'. Preprint.
URL: *https://arxiv.org/abs/1906.09833*

Gravetter, F. J. and Forzano, L.-A. B. (2014), *Research Methods for the Behavioral Sciences*, 5th edn, Cengage Learning, Stamford, CT.

Guerreiro, N. M., Voita, E. and Martins, A. (2023), Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation, *in* 'Proceedings of the 17th Conference of the European Chapter of the

Association for Computational Linguistics', Association for Computational Linguistics, Dubrovnik, Croatia, pp. 1059–1075.
URL: *https://aclanthology.org/2023.eacl-main.75/*

Guillou, L., Hardmeier, C., Lapshinova-Koltunski, E. and Loáiciga, S. (2018), A pronoun test suite evaluation of the English–German MT systems at WMT 2018, *in* O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi and K. Verspoor, eds, 'Proceedings of the Third Conference on Machine Translation: Shared Task Papers', Association for Computational Linguistics, Brussels, Belgium, pp. 570–577.
URL: *https://aclanthology.org/W18-6435/*

Guo, Y. (2020), 'Prospects for the teaching of translation majors in the new era', *Babel* **66**(4-5), 867–881.

Hardman, F. (2020), Embedding a Dialogic Pedagogy in the Classroom: What is the research telling us?, *in* N. Mercer, R. Wegerif and L. Major, eds, 'The Routledge International Handbook of Research on Dialogic Education', Routledge, London, UK, pp. 140–152.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z. and Zhou, M. (2018), 'Achieving Human Parity on Automatic Chinese to English News Translation', *arXiv preprint arXiv:1803.05567* .
URL: *https://arxiv.org/abs/1803.05567*

Hearst, M. A. (2005), Teaching Applied Natural Language Processing: Triumphs and tribulations, *in* 'Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL', Association for Computational Linguistics, pp. 1–8.
URL: *https://aclanthology.org/W05-0101/*

Heer, J. and Bostock, M. (2010), Crowdsourcing Graphical Perception: Using mechanical turk to assess visualization design, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', Association for Computing Machinery, New York, NY, pp. 203–212.

Heesen, R., Bright, L. K. and Zucker, A. (2019), 'Vindicating methodological triangulation', *Synthese* **196**(8), 3067–3081.
URL: *https://doi.org/10.1007/s11229-016-1294-7*

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D. and Pineau, J. (2020), 'Towards the systematic reporting of the energy and carbon footprints of machine learning', *Journal of Machine Learning Research* **21**, 1–43. Paper 20-312, JMLR.

Hmelo-Silver, C. E. (2004), 'Problem-Based Learning: What and how do students learn?', *Educational Psychology Review* **16**(3), 235–266.
URL: *https://doi.org/10.1023/B:EDPR.0000034022.16470.f3*

Honebein, P. C. and Sink, D. L. (2012), 'The Practice of Eclectic Instructional Design', *Performance Improvement* **51**(10), 26–31.
URL: *https://eric.ed.gov/?id=EJ986393*

House, J. (2014), Translation Quality Assessment: Past and Present, *in* J. House, ed., 'Translation: A Multidisciplinary Approach', Palgrave Macmillan, London, pp. 241–264.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. et al. (2025), 'A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions', *ACM Transactions on Information Systems* **43**(2), 1–55.

Huidrom, R. and Belz, A. (2022), A Survey of Recent Error Annotation Schemes for Automatically Generated Text, *in* 'Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)', Association for

Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), pp. 383–398.
URL: *https://aclanthology.org/2022.gem-1.33/*

Hurtado Albir, A. (2007), 'Competence-based curriculum design for training translators', *The Interpreter and Translator Trainer* **1**(2), 163–195.

International Organization for Standardization (2019), 'ISO 9241-210:2019 ergonomics of human-system interaction — part 210: Human-centred design for interactive systems', Standard. ISO/TC 159/SC 4 Ergonomics of Human-System Interaction.

ISO (2015), *Translation services – Requirements for translation services*, International Organization for Standardization, Geneva, Switzerland. ISO 17100:2015.
URL: *https://www.iso.org/standard/59149.html*

ISO (2018), *Risk management — Guidelines*, Geneva, Switzerland.
URL: *https://www.iso.org/standard/65694.html*

Jiang, Y. E., Liu, T., Ma, S., Zhang, D., Yang, J., Huang, H., Sennrich, R., Cotterell, R., Sachan, M. and Zhou, M. (2022), BlonDe: An automatic evaluation metric for document-level machine translation, *in* 'Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', Association for Computational Linguistics, Seattle, United States, pp. 1550–1565.
URL: *https://aclanthology.org/2022.naacl-main.111*

Jin, L., He, J., May, J. and Ma, X. (2023), Challenges in context-aware neural machine translation, *in* 'Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Singapore, pp. 15246–15263.
URL: *https://aclanthology.org/2023.emnlp-main.943/*

Jonassen, D. H., Tessmer, M. and Hannum, W. H. (1999), *Task Analysis Methods for Instructional Design*, Lawrence Erlbaum Associates, Mahwah, NJ.
URL: *https://www.routledge.com/Task-Analysis-Methods-for-Instructional-Design/Jonassen-Tessmer-Hannum/p/book/9780805830866*

Junczys-Dowmunt, M. (2019), Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation, *in* 'Proceedings of the Fourth Conference on Machine Translation (Volume2: Shared Task Papers, Day1)', Association for Computational Linguistics, Florence, Italy, pp. 225–233.
URL: *https://aclanthology.org/W19-5321*

Kenny, D. (2018), Machine Translation, *in* P. Rawling and P. Wilson, eds, 'The Routledge Handbook of Translation and Philosophy', Routledge, London & New York, pp. 428–445.

Kenny, D. and Doherty, S. (2014), 'Statistical Machine Translation in the Translation Curriculum: Overcoming Obstacles and Empowering Translators', *The Interpreter and Translator Trainer* **8**(2), 276–294.

Ker, H. W. (2014), 'Application of Hierarchical Linear Models/Linear Mixed-Effects Models in School Effectiveness Research', *Universal Journal of Educational Research* **2**(2), 173–180.
URL: *https://eric.ed.gov/?id=EJ1053927*

Khatamian Far, P. (2018), 'Challenges of Recruitment and Retention of University Students as Research Participants: Lessons learned from a pilot study', *Journal of the Australian Library and Information Association* **67**(3), 278–292.
URL: *https://doi.org/10.1080/24750158.2018.1500436*

Kim, Y., Tran, D. T. and Ney, H. (2019), When and why is document-level context useful in neural machine translation?, *in* 'Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)', Association for Computa-

tional Linguistics, Hong Kong, China, pp. 24–34.
URL: *https://aclanthology.org/D19-6503*

Klein, D. (2005), A Core-Tools Statistical NLP Course, *in* 'Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL', Association for Computational Linguistics, pp. 23–27.
URL: *https://aclanthology.org/W05-0104/*

Klenowski, V. (2010), Curriculum Evaluation: Approaches and methodologies, *in* 'International Encyclopedia of Education', Elsevier, Amsterdam, Netherlands, pp. 335–341.

Klubička, F. and Kelleher, J. D. (2024), ReproHum #1018-09: Reproducing human evaluations of redundancy errors in data-to-text systems, *in* S. Balloccu, A. Belz, R. Huidrom, E. Reiter, J. Sedoc and C. Thomson, eds, 'Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024', ELRA and ICCL, Torino, Italia, pp. 163–198.
URL: *https://aclanthology.org/2024.humeval-1.16/*

Knowles, R. and Lo, C. (2024), 'Calibration and context in human evaluation of machine translation', *Natural Language Processing* **31**(4), 1017–1041.
URL: *https://www.cambridge.org/core/journals/natural-language-processing/article/calibration-and-context-in-human-evaluation-of-machine-translation/53AD3B45A9E65D5E62DA4F3E825C1BB9*

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Koehn, P., Marie, B., Monz, C., Morishita, M., Murray, K., Nagata, M., Nakazawa, T., Popel, M., Popović, M. and Shmatova, M. (2023), Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet, *in* 'Proceedings of the Eighth Conference on Machine Translation', Association for Computational Linguistics, Singapore, pp. 1–42.
URL: *https://aclanthology.org/2023.wmt-1.1/*

Koehn, P. (2020), *Neural Machine Translation*, Cambridge University Press.

Koehn, P. and Knowles, R. (2017), Six challenges for neural machine translation, *in* 'Proceedings of the First Workshop on Neural Machine Translation', Association for Computational Linguistics, Vancouver, Canada, pp. 28–39.
URL: *https://aclanthology.org/W17-3204/*

Korošec, M. K. (2011), 'Applicability and challenges of using machine translation in translator training', *ELOPE: English Language Overseas Perspectives and Enquiries* **8**(2), 7–18.

Krippendorff, K. (1970), 'Estimating the reliability, systematic error, and random error of interval data', *Educational and Psychological Measurement* **30**(1), 61–70.
URL: *https://journals.sagepub.com/doi/10.1177/001316447003000105*

Krüger, R. (2021), 'AN ONLINE REPOSITORY OF PYTHON RESOURCES FOR TEACHING MACHINE TRANSLATION TO TRANSLATION STUDENTS', *Current Trends in Translation Teaching & Learning E* **8**(4), 4–30.

Krüger, R. (2022), 'Using Jupyter Notebooks as Didactic Instruments in Translation Technology Teaching', *The Interpreter and Translator Trainer* **16**(4), 503–523.

Kühl, N., Goutier, M., Hirt, R. and Satzger, G. (2019), Machine Learning in Artificial Intelligence: Towards a common understanding, Technical report, Technical University of Darmstadt, Department of Business Administration. Working Paper.
URL: *https://arxiv.org/abs/1908.00739*

Kühl, N., Goutier, M., Hirt, R. and Satzger, G. (2020), 'Machine Learning in Artificial Intelligence: Towards a common understanding'. Preprint.
URL: *https://arxiv.org/abs/2004.04686*

Kuhn, T. S. (1962), *The Structure of Scientific Revolutions*, 1st edn, University of Chicago Press, Chicago, IL.

Landis, J. R. and Koch, G. G. (1977), 'The Measurement of Observer Agreement for Categorical Data', *Biometrics* **33**(1), 159–174.
URL: *https://www.jstor.org/stable/2529310*

Lankshear, C. and Knobel, M. (2004), *A Handbook for Teacher Research: From Design to Implementation*, Open University Press, Maidenhead, UK.

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q. and Toral, A. (2020), 'A set of recommendations for assessing human–machine parity in language translation', *Journal of Artificial Intelligence Research* **67**, 653–672.
URL: *https://www.jair.org/index.php/jair/article/view/12164*

Läubli, S., Sennrich, R. and Volk, M. (2018), Has Machine Translation Achieved Human Parity? A Case for Document-Level Evaluation, *in* E. Riloff, D. Chiang, J. Hockenmaier and J. Tsujii, eds, 'Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Brussels, Belgium, pp. 4791–4796.
URL: *https://aclanthology.org/D18-1512*

Lauscher, A., Nozza, D., Crowley, A., Miltersen, E. and Hovy, D. (2023), 'What about *em*? how commercial machine translation fails to handle (neo-)pronouns'. Preprint. Accepted to ACL 2023.
URL: *https://arxiv.org/abs/2305.16051*

Liebel, G. and Chakraborty, S. (2021), 'Ethical Issues in Empirical Studies Using Student Subjects: Re-visiting practices and perceptions', *Empirical Software Engineering* **26**(3), 40.
URL: *https://link.springer.com/article/10.1007/s10664-021-09958-4*

Likert, R. (1932), 'A Technique for the Measurement of Attitudes', *Archives of Psychology* **140**, 1–55.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M. and Zettlemoyer, L. (2020), 'Multilingual denoising pre-training for neural ma-

chine translation', *Transactions of the Association for Computational Linguistics* **8**, 726–742.
URL: *https://aclanthology.org/2020.tacl-1.47/*

Lommel, A., Burchardt, A. and Popović, M. (2014), Assessing inter-annotator agreement for translation error annotation, *in* 'Proceedings of the Workshop on Automatic and Manual Metrics for Operational Translation Evaluation (MTE 2014)', European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 31–37.
URL: *https://mte2014.github.io/MTE2014-Workshop-Proceedings.pdf*

Loock, R. (2020), 'No more rage against the machine: how the corpus-based identification of machine-translationese can lead to student empowerment', *The Journal of specialised translation (JoSTrans)* **34**, 150–170.

Lopes, A., Farajian, M. A., Bawden, R., Zhang, M. and Martins, A. F. T. (2020), Document-level neural MT: A systematic comparison, *in* 'Proceedings of the 22nd Annual Conference of the European Association for Machine Translation', European Association for Machine Translation, Lisboa, Portugal, pp. 225–234.
URL: *https://aclanthology.org/2020.eamt-1.24/*

Luri, L. C. (2016), Kantanlqr: a platform for human evaluation of machine translation output to drive engine rapid improvement, *in* 'Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products', Baltic Journal of Modern Computing, Riga, Latvia.
URL: *https://aclanthology.org/2016.eamt-2.13/*

Luz, S. (2022), Computational Linguistics and Natural Language Processing, *in* 'The Routledge Handbook of Translation and Methodology', Routledge, London, UK, pp. 373–391.

Madureira, B. (2021), Flamingos and Hedgehogs in the Croquet-Ground: Teaching evaluation of NLP systems for undergraduate students, *in* 'Proceedings of the

Fifth Workshop on Teaching NLP', Association for Computational Linguistics, pp. 87–91.
URL: *https://aclanthology.org/2021.teachingnlp-1.12/*

Malpas, J. (2014), Introduction: hermeneutics and philosophy, *in* 'The Routledge companion to hermeneutics', Routledge, pp. 1–10.

Marie, B., Fujita, A. and Rubino, R. (2021), Scientific credibility of machine translation research: A meta-evaluation of 769 papers, *in* 'Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)', Association for Computational Linguistics, Online, pp. 7297–7306.
URL: *https://aclanthology.org/2021.acl-long.566*

Martynova, I., Metelkova, L., Gordeeva, N., Nikitinskaya, L., Emelianova, M. and Trukova, A. (2018), 'THE PROGRAMS OF COMPUTATIONAL LINGUISTICS GRADUATE IN GERMAN AND US UNIVERSITIES', *Visnyk Natsional'noi Akademii Kerivnykh Kadriv Kultury i Mystetstv* **1**(3), 534–540.

Maruf, S., Saleh, F. and Haffari, G. (2021), 'A survey on document-level neural machine translation: Methods and evaluation', *ACM Computing Surveys* **54**(2), 1–36.
URL: *https://doi.org/10.1145/3441691*

Massey, G. and Ehrensberger-Dow, M. (2017), 'Machine learning: Implications for translator education', *Lebende Sprachen* **62**(2), 300–312.

Mathison, S. (1988), 'Why Triangulate?', *Educational Researcher* **17**(2), 13–17.
URL: *https://doi.org/10.3102/0013189X017002013*

Mattoon, J. S. (2005), 'Designing and Developing Technical Curriculum: Finding the right subject matter expert', *Journal of STEM Teacher Education* **42**(2), 5.

Mattos, L. A. d. (1976), *Compendio de Didáctica General*, 3rd edn, Editorial Kapelusz, Buenos Aires, Argentina. Originally published in Portuguese; later translated into Spanish. Used in teacher training courses.

McCrae, J. P. (2024), Co-Creational Teaching of Natural Language Processing, *in* 'Proceedings of the Sixth Workshop on Teaching NLP', Association for Computational Linguistics, Bangkok, Thailand, pp. 33–42.
URL: *https://aclanthology.org/2024.teachingnlp-1.5/*

Melby, A. K. and Foster, C. (2010), 'Context in Translation: Definition, Access, and Teamwork', *Translation & Interpreting* **2**(2), 1–15.
URL: *https://trans-int.org/index.php/transint/article/view/87*

Melero, M., Calvo, B., Rodríguez, M. and Villegas, M. (2023), Language report: Catalan, *in* G. Rehm and A. Way, eds, 'European Language Equality: A Strategic Agenda for Digital Language Equality', Springer Nature, Cham, Switzerland, pp. 107–110.
URL: *https://doi.org/10.1007/978-3-031-28819-7_8*

Mercer, N., Wegerif, R. and Major, L. (2023), *The Routledge International Handbook of Research on Dialogic Education*, Routledge International Handbooks, Routledge, London, UK.

Meyer, J., Craig, A.-M. and Smith, R. (2025), 'Increasing Diversity in STEM Academia: A scoping review of intervention evaluations', *Studies in Higher Education* **50**(3), 456–478.
URL: *https://www.tandfonline.com/doi/full/10.1080/03075079.2024.2442052*

Mitchell-Schuitevoerder, R. (2020), *A Project-Based Approach to Translation Technology*, Translation Practices Explained, Routledge.

Monahan, T. (2002), 'Flexible Space & Built Pedagogy: Emerging IT embodiments', *Inventio* **4**(1), 1–19.
URL: *https://cdr.lib.unc.edu/concern/articles/37720p826*

Moorkens, J. (2018), 'What to Expect from Neural Machine Translation: A Practical In-Class Translation Evaluation Exercise', *The Interpreter and Translator Trainer* **12**(4), 375–387.

Moorkens, J. (2022), Ethics and machine translation, *in* D. Kenny, ed., 'Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence', Vol. 18 of *Translation and Multilingual Natural Language Processing*, Language Science Press, Berlin, Germany, pp. 121–140.
URL: *https://langsci-press.org/catalog/book/342*

Moorkens, J. (2024), ''[I] Am Not a Number': On Quantification and Algorithmic Norms in Translation', *Perspectives* **32**(3), 477–492.

Moorkens, J., Castilho, S., Gaspari, F. and Doherty, S. (2018), Translation quality assessment, *in* J. Moorkens, S. Castilho, F. Gaspari and S. Doherty, eds, 'Translation Quality Assessment: From Principles to Practice', Vol. 1 of *Machine Translation: Technologies and Applications*, Springer International Publishing, Cham, Switzerland, pp. 299–320.

Moorkens, J., Castilho, S., Gaspari, F., Toral, A. and Popović, M. (2024a), 'Proposal for a Triple Bottom Line for Translation Automation and Sustainability: An Editorial Position Paper', *Journal of Specialised Translation* (41), 2–25.
URL: *https://www.jostrans.org/article/view/4706/4263*

Moorkens, J. and Guerberof-Arenas, A. (2024c), Artificial intelligence, automation and the language industry, *in* G. Massey, M. Ehrensberger-Dow and E. Angelone, eds, 'Handbook of the Language Industry: Contexts, Resources and Profiles', Vol. 20 of *Handbooks of Applied Linguistics [HAL]*, De Gruyter, Berlin/Boston, Germany/USA, pp. 71–97.

Moorkens, J., Toral, A., Castilho, S. and Way, A. (2018), 'Translators' Perceptions of Literary Post-editing Using Statistical and Neural Machine Translation', *Translation Spaces* **7**(2), 240–262.

Moorkens, J., Way, A. and Lankford, S. (2024b), *Automating Translation*, Routledge, London.

URL: *https://www.routledge.com/Automating-Translation/Moorkens-Way-Lankford/p/book/9781032436807*

Moorkens, J., Way, A., Lankford, S. et al. (2025), 'Sociotechnical effects of machine translation', *arXiv e-prints* pp. arXiv–2503.

Morgan, D. L. (2007), 'Paradigms Lost and Pragmatism Regained: Methodological implications of combining qualitative and quantitative methods', *Journal of Mixed Methods Research* **1**(1), 48–76.
URL: *https://doi.org/10.1177/2345678906292462*

Morrison, G. R., Ross, S. J., Morrison, J. R. and Kalman, H. K. (2019), *Designing Effective Instruction*, John Wiley & Sons, Hoboken, NJ.

Muhammad, L. N. (2023), 'Guidelines for Repeated Measures Statistical Analysis Approaches with Basic Science Research Considerations', *Journal of Clinical Investigation* **133**(11), e171058.
URL: *https://www.jci.org/articles/view/171058*

Müller, M., Rios Gonzales, M., Sennrich, R. and Scherrer, Y. (2018), A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation, *in* 'Proceedings of the Third Conference on Machine Translation: Research Papers', Association for Computational Linguistics, Brussels, Belgium, pp. 61–72.
URL: *https://aclanthology.org/W18-6307*

Munday, J., Ramos Pinto, S. and Blakesley, J. (2022), *Introducing Translation Studies: Theories and Applications*, 5th edn, Routledge, London & New York.

Munkova, D., Munk, M., Benko, L. and Hajek, P. (2021), 'The role of automated evaluation techniques in online professional translator training', *PeerJ Computer Science* **7**, e706.

Newman, M. and Gough, D. (2020), Systematic Reviews in Educational Research:

Methodology, perspectives and application, *in* 'Systematic Reviews in Educational Research: Methodology, Perspectives and Application', Springer Fachmedien Wiesbaden, pp. 3–22.

Nida, E. A. (1964), *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*, E. J. Brill, Leiden.

Nord, C. (2005), *Text Analysis in Translation: Theory, Methodology, and Didactic Application of a Model for Translation-Oriented Text Analysis*, 2nd edn, Rodopi, Amsterdam & New York.

Nord, C. (2014), *Translating as a Purposeful Activity: Functionalist Approaches Explained*, 1st english ed. edn, Routledge, London & New York.

Nurminen, M. and Koponen, M. (2020), 'Machine Translation and Fair Access to Information', *Translation Spaces* **9**(1), 150–169.
URL: *https://www.benjamins.com/catalog/ts.00025.nur*

Núñez, K. J. (2019), 'Análisis de la percepción, la utilidad y la calidad de los sistemas de TA por parte del traductor en formación', *E-Aesla* **5**, 391–399.

Olkhovska, A. and Frolova, I. (2020), 'Using machine translation engines in the classroom: A survey of translation students' performance', *Advanced Education* **7**(15), 47–55.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002), BLEU: A method for automatic evaluation of Machine Translation, *in* P. Isabelle, E. Charniak and D. Lin, eds, 'Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318.
URL: *https://aclanthology.org/P02-1040/*

Parasuraman, R., Sheridan, T. B. and Wickens, C. D. (2000), 'A Model for Types and Levels of Human Interaction with Automation', *IEEE Transactions on Sys-*

tems, Man, and Cybernetics – Part A: Systems and Humans **30**(3), 286–297.
URL: *https://doi.org/10.1109/3468.844354*

Paulsen Christensen, T., Bundgaard, K., Schjoldager, A. and Dam Jensen, H. (2022), 'What Motor Vehicles and Translation Machines Have in Common - A First Step Towards a Translation Automation Taxonomy', *Perspectives – Studies in Translation Theory and Practice* **30**(1), 19–38.

Perrella, S., Proietti, L., Scirè, A., Barba, E. and Navigli, R. (2024), Guardians of the machine translation meta-evaluation: Sentinel metrics fall in!, *in* 'Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Bangkok, Thailand, pp. 16216–16244.
URL: *https://aclanthology.org/2024.acl-long.856/*

Petticrew, M. and Roberts, H. (2008), *Systematic Reviews in the Social Sciences: A Practical Guide*, revised edition edn, Wiley-Blackwell, Oxford, UK.

Popescu-Belis, A. (2019), 'Context in neural machine translation: A review of models and evaluations', *arXiv preprint arXiv:1901.09115* .

Popović, M. (2015), chrF: character n-gram f-score for automatic mt evaluation, *in* 'Proceedings of the Tenth Workshop on Statistical Machine Translation', Association for Computational Linguistics, Lisbon, Portugal, pp. 392–395.
URL: *https://aclanthology.org/W15-3049/*

Popović, M. (2021), Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output, *in* A. Bisazza and O. Abend, eds, 'Proceedings of the 25th Conference on Computational Natural Language Learning', Association for Computational Linguistics, Online, pp. 234–243.
URL: *https://aclanthology.org/2021.conll-1.18/*

Post, M. and Lopez, A. (2014), 'The machine translation leaderboard', *The Prague Bulletin of Mathematical Linguistics* **102**(1), 37–46.

Prates, M. O., Avelar, P. H. and Lamb, L. C. (2020), 'Assessing gender bias in machine translation: a case study with google translate', *Neural Computing and Applications* **32**, 6363–6381.

Pym, A. (2023), *Exploring Translation Theories*, Routledge Translation Studies Series, Routledge, London & New York.

Ramírez-Sánchez, G. (2023), MutNMT, an open-source nmt tool for educational purposes, *in* 'Proceedings of the 1st Workshop on Open Community-Driven Machine Translation', European Association for Machine Translation, Tampere, Finland, p. 31.
URL: *https://aclanthology.org/2023.crowdmt-1.5*

Ranathunga, S., Lee, E. A., PriftiSkenduli, M., Shekhar, R., Alam, M. and Kaur, R. (2023), 'Neural machine translation for low-resource languages: A survey', *ACM Computing Surveys* **55**(11), 1–37.
URL: *https://dl.acm.org/doi/10.1145/3567592*

Rands, M. L. and Gansemer-Topf, A. M. (2017), 'The Room Itself Is Active: How classroom design impacts student engagement', *Journal of Learning Spaces* **6**(1), 26–33.
URL: *https://files.eric.ed.gov/fulltext/EJ1152568.pdf*

Raunak, V., Kocmi, T. and Post, M. (2024), SLIDE Reference-free evaluation for machine translation using a sliding document window, *in* K. Duh, H. Gomez and S. Bethard, eds, 'Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)', Association for Computational Linguistics, Mexico City, Mexico, pp. 205–211.
URL: *https://aclanthology.org/2024.naacl-short.18/*

Raunak, V., Menezes, A. and Junczys-Dowmunt, M. (2021), The curious case of hallucinations in Neural Machine Translation, *in* 'Proceedings of the 2021 Conference

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', Association for Computational Linguistics, Online, pp. 1172–1183.
URL: *https://aclanthology.org/2021.naacl-main.92/*

Rei, R., Stewart, C., Farinha, A. C. and Lavie, A. (2020), COMET: A neural framework for MT evaluation, *in* B. Webber, T. Cohn, Y. He and Y. Liu, eds, 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Online, pp. 2685–2702.
URL: *https://aclanthology.org/2020.emnlp-main.213/*

Reijers, W. and Dupont, Q. (2023), Prolegomenon to contemporary ethics of machine translation, *in* H. Moniz and C. Parra Escartín, eds, 'Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation', Vol. 18 of *Machine Translation: Technologies and Applications*, Springer International Publishing, Cham, Switzerland, pp. 11–27.
URL: *https://doi.org/10.1007/978-3-031-14689-3_2*

Rico, C. and González Pastor, D. (2022), 'The role of machine translation in translation education: a thematic analysis of translator educators' beliefs', *Translation & Interpreting: The International Journal of Translation and Interpreting Research* **14**(1), 177–197.

Rodríguez de Céspedes, B. (2019), 'Translator education at a crossroads: the impact of automation', *Lebende Sprachen* **64**(1), 103–121.

Rossi, C. (2017), 'Introducing statistical machine translation in translator training: From uses and perceptions to course design, and back again', *Revista Tradumàtica: Tecnologies de la Traducció* **1**(15), 48–62.

Rossman, G. B. and Wilson, B. L. (1985), 'Numbers and Words: Combining quantitative and qualitative methods in a single large-scale evaluation study', *Evaluation*

Review **9**(5), 627–643.

URL: *https://doi.org/10.1177/0193841X8500900505*

Rothwell, A. and Svoboda, T. (2019), 'Tracking translator training in tools and technologies: findings of the emt survey 2017', *Journal of Specialised Translation* **32**, 26.

Rysova, K., Rysova, M., Musil, T., Polakova, L. and Bojar, O. (2019), 'A test suite and manual evaluation of document-level nmt at wmt19', *arXiv preprint arXiv:1908.03043* .

Saçak, B., Bozkurt, A. and Wagner, D. A. (2022), 'Learning Design versus Instructional Design: A bibliometric study through data visualization approaches', *Education Sciences* **12**(11), 752.

URL: *https://doi.org/10.3390/educsci12110752*

Saldanha, G. and O'Brien, S. (2014), *Research Methodologies in Translation Studies*, Routledge, London, UK.

Saunders, D., Stahlberg, F. and Byrne, B. (2020), Using context in neural machine translation training objectives, *in* D. Jurafsky, J. Chai, N. Schluter and J. Tetreault, eds, 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 7764–7770.

URL: *https://aclanthology.org/2020.acl-main.693/*

Schnell, B. and Rodríguez, N. (2017), 'Ivory tower vs. workplace reality: Employability and the t&i curriculum — balancing academic education and vocational requirements: A study from the employers' perspective', *The Interpreter and Translator Trainer* **11**(2–3), 160–186.

Seburn, T. (2021), *How to Write Inclusive Materials*, ELT Teacher 2 Writer.

URL: *https://eltteacher2writer.co.uk/our-books/how-to-write-inclusive-materials/*

Seel, N. M. and Dijkstra, S. (2004), *Curriculum, Plans, and Processes in Instructional Design: International Perspectives*, Routledge, London, UK.

Seel, N. M., Lehmann, T., Blumschein, P. and Podolskiy, O. A. (2017), *Instructional Design for Learning: Theoretical Foundations*, Springer, Cham, Switzerland.

Sellam, T., Das, D. and Parikh, A. P. (2020), BLEURT: Learning robust metrics for text generation, *in* D. Jurafsky, J. Chai, N. Schluter and J. Tetreault, eds, 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 7881–7892.
URL: *https://aclanthology.org/2020.acl-main.704/*

Sheng, E., Natarajan, P., Gordon, J. and Burns, G. (2017), An Investigation into the Pedagogical Features of Documents, *in* 'Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications', Association for Computational Linguistics, Copenhagen, Denmark, pp. 109–120.
URL: *https://aclanthology.org/W17-5012/*

Shterionov, D. and Vanmassenhove, E. (2023), The ecological footprint of neural machine translation systems, *in* 'Towards Responsible Machine Translation', Translation and Sustainability, Springer Nature, Cham, Switzerland, pp. 185–213.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006), A study of translation edit rate with targeted human annotation, *in* 'Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers', Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA, pp. 223–231.
URL: *https://aclanthology.org/2006.amta-papers.25/*

Steiner, G. (1998), *After Babel: Aspects of Language and Translation*, Oxford University Press.

Strubell, E., Ganesh, A. and McCallum, A. (2019), Energy and policy considerations for deep learning in NLP, *in* A. Korhonen, D. Traum and L. Màrquez, eds,

'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 3645–3650.

URL: *https://aclanthology.org/P19-1355/*

Suojanen, T., Koskinen, K. and Tuominen, T. (2014), *User-Centered Translation*, Translation Practices Explained, Routledge, London & New York.

Tashakkori, A. and Teddlie, C. (1998), *Mixed Methodology: Combining Qualitative and Quantitative Approaches*, SAGE Publications, Thousand Oaks, CA.

TAUS (2019), 'Dynamic quality framework'. Accessed: 2024-06-10.

URL: *https://www.taus.net/quality-evaluation/dynamic-quality-framework*

Terribile, S. (2024), Productivity in the Post-Editing of Neural Machine Translation: A Mixed-Methods Analysis of Speed and Edits at Toppan Digital Language, PhD thesis, The University of Manchester.

Tiedemann, J. (2012), Parallel data, tools and interfaces in OPUS, *in* 'Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)', European Language Resources Association (ELRA), Istanbul, Turkey, pp. 2214–2218.

URL: *http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf*

Tiedemann, J. and Scherrer, Y. (2017), Neural Machine Translation with Extended Context, *in* 'Proceedings of the Third Workshop on Discourse in Machine Translation', Association for Computational Linguistics, Copenhagen, Denmark, pp. 82–92.

URL: *https://aclanthology.org/W17-4811*

Toral, A., Castilho, S., Hu, K. and Way, A. (2018), Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation, *in* 'Proceedings of the Third Conference on Machine Translation: Research Papers', Association

for Computational Linguistics, Brussels, Belgium, pp. 113–123.
URL: *https://aclanthology.org/W18-6312/*

Tronchoni, H., Izquierdo, C. and Anguera, M. T. (2022), 'A Systematic Review on Lecturing in Contemporary University Teaching', *Frontiers in Psychology* **13**, 971617.
URL: *https://www.frontiersin.org/articles/10.3389/fpsyg.2022.971617/full*

Tsujii, J. (2011), Computational Linguistics and Natural Language Processing, *in* 'Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing', Vol. Part I, 6608 of *Lecture Notes in Computer Science*, Springer, pp. 52–67.

Uszkoreit, H., Kordoni, V., Kubon, V., Rosner, M. and Kirchmeier-Andersen, S. (2005), Language Technology from a European Perspective, *in* 'Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL', Association for Computational Linguistics, pp. 43–48.
URL: *https://aclanthology.org/W05-0106/*

Vagia, M., Transeth, A. A. and Fjerdingen, S. A. (2016), 'A Literature Review on the Levels of Automation during the Years: What are the different taxonomies that have been proposed?', *Applied Ergonomics* **53**, 190–202.
URL: *https://doi.org/10.1016/j.apergo.2015.09.013*

van Rijsbergen, C. J. (1979), *Information Retrieval*, 2nd edn, Butterworth-Heinemann, London, UK.

Vanroy, B., Tezcan, A. and Macken, L. (2023), MATEO: MAchine Translation Evaluation Online, *in* M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. Alvarez-Vidal, N. Aranberri, M. Nunziatini, C. Parra-Escartín, M. Forcada, M. Popović, C. Scarton and H. Moniz, eds, 'Proceedings of the 24th Annual Conference of the European Association for Machine Translation', European Association for Machine

Translation, Tampere, Finland, pp. 499–500.
URL: *https://aclanthology.org/2023.eamt-1.52/*

Vattimo, G. (2015), Conclusion: The Future of Hermeneutics, *in* J. Malpas and H.-H. Gander, eds, 'The Routledge Companion to Hermeneutics', Routledge, London & New York, pp. 721–728.

Venkatesan, H. (2018), 'Teaching translation in the age of neural machine translation', *APLX 2017 at Taipei Tech-Transformation and Development: Language, Culture, Pedagogy and Translation* pp. 39–54.

Venuti, L. (2019), *Contra Instrumentalism: A Translation Polemic*, Provocations, University of Nebraska Press, Lincoln.

Vernikos, G., Thompson, B., Mathur, P. and Federico, M. (2022), Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric, *in* P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-Jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzmán, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi and M. Zampieri, eds, 'Proceedings of the Seventh Conference on Machine Translation (WMT)', Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), pp. 118–128.
URL: *https://aclanthology.org/2022.wmt-1.6/*

Vieira, L. N., O'Hagan, M. and O'Sullivan, C. (2020), 'Understanding the Societal Impacts of Machine Translation: A Critical Review of the Literature on Medical and Legal Use Cases', *Information, Communication & Society* **24**(11), 1515–1532.

Voita, E., Sennrich, R. and Titov, I. (2019), When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Flo-

rence, Italy, pp. 1198–1212.
URL: *https://aclanthology.org/P19-1116*

Walker, E. A., Redfern, A. and Oleson, J. J. (2019), 'Linear Mixed-Model Analysis to Examine Longitudinal Trajectories in Vocabulary Depth and Breadth in Children Who Are Hard of Hearing', *Journal of Speech, Language, and Hearing Research* **62**(3), 525–542.
URL: *https://pubs.asha.org/doi/10.1044/2018_JSLHR-L-ASTM-18-0250*

Wang, E. L., Kaufman, J. H., Lee, S., Kim, B. and Opfer, V. D. (2023), 'Instructional System Coherence: A scoping literature review'. Accessed February 21, 2025.
URL: *https://www.rand.org/pubs/research_reports/RRA279-5.html*

Way, A. (2018), Quality expectations of machine translation, *in* J. Moorkens, S. Castilho, F. Gaspari and S. Doherty, eds, 'Translation Quality Assessment: From Principles to Practice', Springer International Publishing, Cham, Switzerland, pp. 159–178.
URL: *https://doi.org/10.1007/978-3-319-91241-7_8*

Wilks, Y. (1978), Machine translation and artificial intelligence Implementing machine aids to translation, *in* B. M. Snell, ed., 'Translating and the Computer', Aslib Proceedings, London, UK.
URL: *https://aclanthology.org/1978.tc-1.2/*

Wong, B. T. M. and Kit, C. (2012), Extending machine translation evaluation metrics with lexical cohesion to document level, *in* J. Tsujii, J. Henderson and M. Paşca, eds, 'Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning', Association for Computational Linguistics, Jeju Island, Korea, pp. 1060–1068.
URL: *https://aclanthology.org/D12-1097/*

Woods, A., Luke, A. and Weir, K. (2010), Curriculum and Syllabus Design, *in* 'In-

ternational Encyclopedia of Education, 3rd Edition, Volume 1', Elsevier, Oxford, UK, pp. 362–367.

Xiong, H., He, Z., Wu, H. and Wang, H. (2019), Modeling coherence for discourse neural machine translation, *in* 'Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence', AAAI Press, Honolulu, Hawaii, USA, pp. 7338–7345. **URL:** *https://aaai.org/ojs/index.php/AAAI/article/view/4721/*

Xu, M. and Wang, C. (2011), 'Translation students' use and evaluation of online resources for chinese-english translation at the word level', *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association* **6**(1), 62–86.

Zawacki-Richter, O., Kerres, M., Bedenlier, S., Bond, M. and Buntins, K., eds (2020), *Systematic Reviews in Educational Research: Methodology, Perspectives and Application*, Springer VS / Springer Nature, Wiesbaden, Germany.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. and Artzi, Y. (2019), 'BERTScore: Evaluating Text Generation with BERT', *arXiv preprint arXiv:1904.09675* . **URL:** *https://arxiv.org/abs/1904.09675*

Zhao, W., Glavaš, G., Peyrard, M., Gao, Y., West, R. and Eger, S. (2020), On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation, *in* D. Jurafsky, J. Chai, N. Schluter and J. Tetreault, eds, 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 1656–1671. **URL:** *https://aclanthology.org/2020.acl-main.151/*

Zhao, W., Strube, M. and Eger, S. (2023), DiscoScore: Evaluating text generation with BERT and discourse coherence, *in* H. Bouamor, J. Pino and K. Bali, eds, 'Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics', Association for Computational Linguistics,

Dubrovnik, Croatia, pp. 3865–3883.

URL: *https://aclanthology.org/2023.eacl-main.278/*

Zierer, K. and Seel, N. M. (2012), 'General Didactics and Instructional Design: Eyes like twins — a transatlantic dialogue about similarities and differences, about the past and the future of two sciences of learning and teaching', *SpringerPlus* **1**(1), 15.

URL: *https://springerplus.springeropen.com/articles/10.1186/2193-1801-1-15*

Appendix A

Systematic Literature Review Protocol

Title

A Systematic Review on Teaching Translation Quality Assessment

Researchers

- João Lucas Cavalheiro Camargo (Researcher);
- Dr. Sheila Castilho (Supervisor 1)
- Dr. Joss Moorkens (Supervisor 2)

Description

This systematic review is part of a literature review. It aims to scope the literature on machine translation quality assessment (MTQA) teaching in Translation Studies (TS) and Natural Language Processing (NLP).

Objective

Identify and analyse teaching materials, procedures, metrics, methods, etc. on MTQA education.

Research Questions

- SRL- Q1 - How is MTQA taught in TS education?
- SRL- Q2 – How is MTQA taught in NLP education?

- SRL- Q3 - How is MTQA taught in different teaching modalities in TS and NLP education?
- SRL-Q4 - What are the rationales for MTQA in TS and NLP education?
- SRL-Q5 - What educational materials are used for MTQA teaching in TS and NLP education?

Studies Identification

Keywords

- **English:** Translation Teaching; Translation Quality Assessment; Translation Pedagogy; Translation Didactics; Translator Training; Translator Education; Machine Translation Evaluation Selection and Evaluation of Studies Inclusion and Exclusion Criteria
- **Portuguese:** Ensino de Tradução; Avaliação de Qualidade de Tradução; Pedagogia da Tradução; Didática da Tradução; Formação de Tradutores; Avaliação de Tradução Automática
- **Spanish:** Enseñanza de la traducción; Evaluación de la calidad de la traducción; Pedagogía de la traducción; Didáctica de la traducción; Formación de traductores; Evaluación de traducción automática

Search Strings Strategy

The search strings were created and implemented with different operators to obtain exact, limited or expanded results. Exact results were retrieved using inverted commas around the search term. Operators such as **OR** were used to broaden results, while **AND** was used to combine results. Parentheses or brackets were employed to group multiple terms when databases allowed.

Databases list and search strings

Stage 1 - Translation Studies

1. Institute of Education Sciences (https://eric.ed.gov/) **Search string:** ("machine translation" OR "automatic translation" OR "tradução automática" OR

"human translation" OR "tradução humana" OR "traducción automática" OR "traducción humana") AND (quality OR evaluat* OR avaliação OR qualidade OR evaluación OR calidad) AND (teaching OR pedagogy OR didactics OR education OR ensino OR formação OR pedagogia OR didática OR didáctica OR pedagogía OR formación)

Results: 842 entries

After removing duplicates: 821

Date of Access: 28/11/2021

After Inclusion/Exclusion Criteria:

MTQA for Translation Studies - 20 entries

2. Hispanic American Periodical Index (HAPI) (https://hapi.ucla.edu/) No results.

   Date of Access: 28/11/2021

3. JSTOR (https://www.jstor.org/)

   Search string: ((((”machine translation”) OR (”human translation”)) AND (”evaluation”)) AND (”teaching”))

   Results: 76 entries

   After removing duplicates: 64

   Date of Access: 05/12/2021

   After inclusion/exclusion criteria:

   MTQA for Translation Studies - 2

4. MLA International Bibliography (https://www.mla.org/Publications/MLA-International-Bibliograph)

   Search String: machine translation OR human translation AND teaching AND quality assessment AND evaluation

Results: 521 entries

Exported: 364 entries

Date of access: 25/01/2022

After removing duplicates: 358

After inclusion/exclusion criteria:

MTQA for Translation Studies - 24

5. Science direct: (https://www.sciencedirect.com/)

   Search string: ("machine translation" OR "human translation") AND ("quality" OR "assessment" OR "evaluation") AND ("teaching" OR "pedagogy")
   Year: 2011-2021

   Results: 279 entries

   Date of access: 03/02/2021

   After inclusion/exclusion criteria:

   MTQA for Translation Studies - 3

6. Taylor & Francis Online (https://www.tandfonline.com/)

   Search string: [[All: "machine translation"] OR [All: "automatic translation"] OR [All: "tradução automática"] OR [All: "human translation"] OR [All: "tradução humana"] OR [All: "traducción automática"] OR [All: "traducción humana"]] AND [[All: quality] OR [All: evaluat*] OR [All: avaliação] OR [All: qualidade] OR [All: evaluación] OR [All: calidad]] AND [[All: teaching] OR [All: pedagogy] OR [All: didactics] OR [All: education] OR [All: ensino] OR [All: formação] OR [All: pedagogia] OR [All: didática] OR [All: didáctica] OR [All: pedagogía] OR [All: formación]] AND [Publication Date: (01/01/2011 TO 12/31/2021)]

   Results: 481

   Date of access: 04/02/2022

After removing duplicates: 477

After inclusion/exclusion criteria:

MTQA for Translation Studies - 39

7. Web of Science (https://www.webofscience.com/)

   Search string: ALL=((“machine translation” OR “automatic translation” OR “tradução automática” OR “human translation” OR “tradução humana” OR “traducción automática” OR “traducción humana”) AND (quality OR evaluat* OR avaliação OR qualidade OR evaluación OR calidad) AND (teaching OR pedagogy OR didactics OR education OR ensino OR formação OR pedagogia OR didática OR didáctica OR pedagogía OR formación))

   Results: 344

   Date of access: 14/02/2022

   After inclusion/exclusion criteria **MTQA for Translation Studies** - 15

8. Wiley Online Library (https://onlinelibrary-wiley-com)

   Search string: ”(“machine translation” OR “automatic translation” OR “tradução automática” OR “human translation” OR “tradução humana” OR “traducción automática” OR “traducción humana”) AND (quality OR evaluat* OR avaliação OR qualidade OR evaluación OR calidad) AND (teaching OR pedagogy OR didactics OR education OR ensino OR formação OR pedagogia OR didática OR didáctica OR pedagogía OR formación)”

   Results: 509

   Date of access: 15/02/2022

   After removing duplicates: 167

   After inclusion/exclusion criteria:

   MTQA for Translation Studies - 3

9. Portal de Periódicos da Capes (http://novo.periodicos.capes.gov.br/)

Search string: “formação de tradutores” AND “ensino de tradução” AND “avaliação” AND “qualidade” OR “enseñanza de la traducción” OR “pedagogía de la traducción” OR “evaluación de traducción”

Results: 137 entries

Date of Access: 16/02

After inclusion/exclusion criteria:

MTQA for Translation studies - 6

10. Scientific Electronic Library Online (www.scielo.org)

    Search string 1: (”machine translation” OR ”human translation” OR ”traducción automática” OR ”traducción humana” OR ”tradução automática” OR ”tradução humana”) AND (”teaching” OR ”pedagogy” OR ”didactics” OR ”education” OR ”ensino” OR ”formação” OR ”pedagogia” OR ”didática” OR ”didáctica” OR ”pedagogía” OR ”formación”)

    Results: 5

    Search string 2: (“translation quality assessment” OR ”avaliação da qualidade da tradução” OR “evaluación de calidad de tradução”

    Results: 2

    Date of access: 16/02/2022

    After inclusion/exclusion process: 5

    MTQA for Translation Studies: 4

11. SAGE Journals (https://journals.sagepub.com/)

    Search String: 1 - [All ”machine translation”] AND [All evaluation] AND [All teaching] AND [All quality]

    Results: 48 entries

    Date of access: 21/02/2022

2 - [All "human translation"] AND [All evaluation] AND [All teaching] AND [All quality]

Results: 11 entries

Date of access: 21/02/2022

3 - [All machine translation] AND [All human translation] AND [All "teaching"] AND [All "evaluation"] AND [All "quality assessment"]

Results: 112 entries

Date of access: 21/02/2022

After inclusion/exclusion selection criteria - 1 result

MTQA for Translation Studies - 1

12. Google Scholar OPTION 3 (selected)

Results: 251 results

Strings: "neural machine translation" AND "statistical machine translation" AND "human translation" AND "evaluation" AND "quality" AND "teaching"

Date of Access: 08/02/2022

OPTION 4 (selected)

Results: 33

Strings: "tradução automática" AND "tradução humana" AND "formação de tradutores" AND "qualidade" AND "avaliação"

Date of Access: 23/02/2022

OPTION 5 (selected)

Results: 79

Strings: "traducción automática" AND "traducción humana" AND "calidad" AND "evaluación" AND "enseñanza" AND "formación"

Date of Access: 23/02/2022

After inclusion/exclusion process:

MTQA for Translation Studies - 28 studies

Stage 2 - Natural Language Processing

1. Institute of Education Sciences

   Search string: ("natural language processing students") AND ("machine translation") AND (evaluat*) AND (course OR module OR teaching) Filters: Last 10 years; Higher Education

   Results: 142 studies

   Date of Access: 04/08

   After Inclusion/Exclusion criteria: 0 studies

2. HAPI **Search string:** "natural language processing" OR "machine translation" AND "evaluation" AND "teaching"

   Results: 3 studies

   Date of Access: 17/08

   After Inclusion/Exclusion criteria: 0 studies

3. JSTOR

   Search string 1: "natural language processing" AND "machine translation" AND "teaching"

   Results: 50 studies

   Date of Access: 04/08

   After Inclusion/Exclusion criteria: 0 studies

   Search string 2: "natural language processing" AND "course" AND "machine translation" AND "evaluation" AND "teaching"

   Results: 29 studies

   Date of Access: 04/08

After Inclusion/Exclusion criteria: 0 studies

4. MLA International Bibliography

   Search string: “natural language processing” AND “machine translation” AND “teaching”

   Results: 2

   Date of Access: 04/08

   After Inclusion/Exclusion criteria: 0 studies

5. Science Direct

   Search string: (”natural language processing”) AND ”teaching” AND (”machine translation”) AND (course OR module)

   Results: 137

   Date of Access: 21/08

   After Inclusion/Exclusion criteria: 4 studies

6. Taylor & Francis

   Search string: [All: ”natural language processing”] AND [All: ”teaching”] AND [All: ”machine translation”] AND [[All: course] OR [All: module]] AND [Publication Date: (01/01/2011 TO 12/31/2023)

   Results: 41 studies

   Date of Access: 21/08

   After Inclusion/Exclusion criteria: 0 studies

7. Web of Science

   Search string: “machine translation” AND “nlp” AND “teaching” AND “education” AND “curriculum” Filters: natural language processing/ teaching/ computer science/ natural language processing student/ Timespan: 2011-01-01 to 2023-12-31

Results: 120 studies

Date of Access: 21/08

After Inclusion/Exclusion criteria: 1 study

Stage 1 - Translation Studies The criteria for inclusion of studies are the following:

- 1 - Studies must be related to Translation Studies, above all, in Translator Education, Translation Teaching, Translation Didactics, Translation Pedagogy, Machine Translation and Machine Translation Evaluation;
- 2 - Studies must be published and I must have full access to them;
- 3 - They must describe or present the methodology in the study;
- 4 - They must be peer reviewed;
- 5 - Studies must present methods, education modalities, methodologies and/or teaching approaches.
- 6 - Studies must have been published between 2011 to 2021.

The criteria for exclusion of studies are the following:

- 1 - Studies not fully available;
- 2 - Studies published before 2011;
- 3 - Duplicate entries;
- 4 - Studies without an abstract;
- 5 - Studies in languages other than Portuguese, Spanish and English;
- 6 - Studies not related to MT evaluation, such as non-MT evaluation methods;

- 7 - Studies in MT but unrelated to MT evaluation, such as post-editing* or MT literacy.

  *: This refers to the training of post-editing itself, not using post-editing as an evaluation approach.

Stage 2 - Natural Language Processing The criteria for the inclusion of the studies are the following:

- 1 - Studies must be related to Natural Language Process Teaching, specifically, Machine Translation Evaluation teaching for NLP audiences
- 2 - Studies must be published and I must have full access to them;
- 3 - They must describe or present the content of the MT evaluation course/module in the study;
- 4 - They must be peer reviewed;
- 5 - Studies must present MT evaluation as a topic or content in the outline of syllabi, modules or workshop descriptions.
- 6 - Studies must have been published between 2011 to 2023.

The criteria for the exclusion of studies are the following:

- 1 - Studies not fully available;
- 2 - Studies published before 2011;
- 3 - Duplicate entries;
- 4 - Studies without an abstract;
- 5 - Studies in languages other than Portuguese, Spanish and English;
- 6 - Studies about NLP applied to education in other areas of study other than MT.

Strategy for Selection of Studies Studies will be selected through the following criteria, in different levels of priority, in this order:

- Reading hte Title and Abstract
- Reading the Keywords
- Reading the Methods

Evaluation of Studies Studies will be read in this order after selection: Title - Abstract - Keywords - Conclusions - Methods and Analyses (or equivalents)

Appendix B

Systematic Literature Review

Selection

Phase 1 - Translation Studies					
AUTHOR(S)	#	TITLE	YEAR	TYPE OF DOCUMENT	CODE
Koletnik Korošec, Melita	1	Applicability and Challenges of Using Machine Translation in Translator Training	2011	Journal Article	RTL, CSL
Wan, Xiaojun		Bilingual Co-Training for Sentiment Classification of Chinese Product Reviews	2011	Journal Article	Rejected
Xu, Mianjun / Wang, Caiwen	2	Translation Students' Use and Evaluation of Online Resources for Chinese-English Translation at the Word Level	2011	Journal Article	RTL, MOD

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Flórez, Silvia / Alcina, Amparo	3	'Free/Open-Source Software for the Translation Classroom'	2011	Journal Article	RE
Krüger, Ralph		Working with Corpora in the Translation Classroom	2012	Journal Article	Rejected
Cantarotti, Aline / Lourenco, Fernanda Maria Alves		Translation–an approach for teaching/learning for the Executive Secretary/Traducao– uma abordagem de ensino/aprendizagem para o Secretariado Executivo	2012	Journal Article	Rejected
Dejica-Cartis, Daniel	4	Developing the Electronic Tools for Translators Syllabus at Politehnica University of Timisoara	2012	Journal Article	CSL
Beaven, Tita / Comas-Quinn, Anna / Hauck, Mirjam / de los Arcos, Beatriz / Lewis, Timothy		The Open Translation MOOC: Creating Online Communities to Transcend Linguistic Barriers	2013	Journal Article	Rejected
Takewa, Mika		Use of Discussion Board and PaperShow in Translation Class	2013	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Murphy, Amanda C		Incorporating Editing into the Training of English Language Students in the Era of English as a Lingua Franca	2013	Journal Article	Rejected.
Plaza Lara, Cristina		La enseñanza de la traducción asistida por ordenador: análisis didáctico-curricular y propuesta de integración de la subcompetencia instrumental-profesional en el aula	2014	PhD Thesis	Rejected
Valor Miró, J. D. / Spencer, R. N. / Pérez González de Martos, A. / Garcés Díaz-Munío, G. / Turró, C. / Civera, J. / Juan, A.		Evaluating Intelligent Interfaces for Post-Editing Automatic Transcriptions of Online Video Lectures	2014	Journal Article	Excluded
Dolynskyi, Yevhen		Peculiarities of Formation the Informative Competence of Future Translators in EU Universities	2014	Journal Article	Excluded
Post, Matt / Lopez, Adam	5	The Machine Translation Leaderboard	2014	Journal Article	RE

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Doherty, Stephen / Kenny, Dorothy	6	The design and evaluation of a Statistical Machine Translation syllabus for translation students	2014	Journal Article	RE, TPM, CSL, MOD
Kenny, Dorothy / Doherty, Stephen	7	Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators	2014	Journal Article	RTL
Flanagan, Marian / Christensen, Tina Paulsen		Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes	2014	Journal Article	Rejected
Robert, Isabelle S. / Waes, Luuk Van		Selecting a translation revision procedure: do common sense and statistics agree?	2014	Journal Article	Excluded
Espunya, Anna		The UPF learner translation corpus as a resource for translator training	2014	Journal Article	Rejected
Aqeeli, Mohammed Mohammed al / Jadhav, B. S.		Potentials of Post-MT Editing as a Teaching Method to Promote Translation Quality	Time Spent	2015	Journal Article & Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Vieira, Teresinha Penaforte / Filho, Pedro Henrique Lima Praxedes		Bilingual sub-competence and meta-reflection: a pedagogical intervention with translators-in-training on aspects of texture via Systemic-Functional Linguistics/ Subcompetencia bilingue e metarreflexao: uma intervencao pedagogica com tradutores em formacao sobre aspectos da tessitura textual via Linguistica Sistemico-Funcional	2016	Journal Article	Rejected
Salmi, Leena / Kinnunen, Tuija		Training translators for accreditation in Finland	2015	Journal Article	Rejected
González, Marta García / Díaz, María Teresa Veiga		Guided Inquiry and Project-Based Learning in the field of specialised translation: a description of two learning experiences	2015	Journal Article	Rejected
Rodríguez-Castro, Mónica / Sullivan, Clare E		Rethinking the legal translation classroom: a course for legal translation professionals	2015	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Yamada, Masaru		Can college students be post-editors? An investigation into employing language learners in machine translation plus post-editing settings	2015	Journal Article	Rejected
Gaspari, Federico / Almaghout, Hala / Doherty, Stephen		A Survey of Machine Translation Competences: Insights for Translation Technology Educators and Practitioners	2015	Journal Article	Rejected
Schmidhofer, Astrid		La auto-revisión como herramienta en la enseñanza de la traducción	2016	Thesis	Rejected
Araghizade, Elmira / Jadidi, Esmaeil		The Impact of Translators' Epistemological Beliefs and Gender on Their Translation Quality	2016	Journal Article	Rejected
Rafieyan, Vahid		Effect of Cultural Distance on Translation of Culture-Bound Texts	2016	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Serpa, Talita / Rocha, Celso Fernando		Corpus based translation pedagogy: theory and practice for teaching translational habitus using corpora/Pedagogia da tradução baseada em corpus: associações entre teorias e práticas do ensino do habitus tradutório por meio de corpora	2016	Journal Article	Rejected
Corazza, Sandra Mara		Currículo e Didática da Tradução: vontade, criação e crítica	2016	Journal Article	Rejected
Ivanova, Oksana	8	Translation and ICT Competence in the Globalized World	2016	Journal Article	RTL
Tomozeiu, Daniel / Koskinen, Kaisa / D'Arcangelo, Adele		Teaching intercultural competence in translator training	2016	Journal Article	Rejected
Farrell, Michael	9	Building a Custom Machine Translation Engine as part of a Postgraduate University Course: a Case Study	2017	Conference Paper	TPM, CSL

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Carmo, Félix Emanuel Martins do		Post-editing: a theoretical and practical challenge for translation studies and machine learning	2017	PhD Dissertation	Rejected
Marín Juarros, Victoria		La formación de poseditores mediante el uso de las TIC: estado de la cuestión y una propuesta de estrategia de formación	2017	Master's Thesis	Rejected
Insai, Sakolkarn / Poonlarp, Tongtip		More Heads Are Better than One: Peer Editing in a Translation Classroom of EFL Learners	2017	Journal Article	Rejected
Olkhovska, Alla		Testing Efficiency of the Methodology of Teaching Students Majoring in Philology to Translate Texts Using Cat-Tools: A Pilot Study	2017	Journal Article	Rejected
Massey, Gary / Ehrensberger-Dow, Maureen	10	Machine Learning: Implications for Translator Education	2017	Journal Article	RTL
Liparini, Tania / Leipnitz, Luciane		Competência Tradutória: o desenvolvimento da subcompetência sobre conhecimentos em tradução	2017	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
LIPARINI CAMPOS, Tania / LEIPNITZ, Luciane / BRAGA, Camila N. de Oliveira		Avaliação da Qualidade da Tradução: resultados da primeira fase de um estudo longitudinal sobre a aquisição da competência tradutória	2017	Journal Article	Rejected
Mellinger, Christopher D.		Translators and machine translation: knowledge and skills gaps in translator pedagogy	2017	Journal Article	Rejected
Bywood, Lindsay / Georgakopoulou, Panayota / Etchegoyhen, Thierry		Embracing the threat: machine translation as a solution for subtitling	2017	Journal Article	Rejected
Schnell, Bettina / Rodríguez, Nadia	11	Ivory tower vs. workplace reality	2017	Journal Article	RTL, CSL
Bowker, Lynne / McBride, Cheryl		Précis-writing as a form of speed training for translation students	2017	Journal Article	Rejected
Xiao, Rui / Yu, Huan		Global trends in translator and interpreter training: mediation and culture	2017	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Al-Batineh, Mohammed / Bilali, Loubna		Translator training in the Arab world: are curricula aligned with the language industry?	2017	Journal Article	Rejected
Robert, Isabelle S. / Remael, Aline / Ureel, Jim J. J.		Towards a model of translation revision competence	2017	Journal Article	Rejected
Rossi, Caroline	12	Introducing statistical machine translation in translator training: from uses and perceptions to course design, and back again	2017	Journal Article	RTL, CSL
Lee, Jaerang		Designing Computer-Assisted Translation Software for Trainee Translators Based on Classroom Parallel Corpora	2018	Journal Article	Rejected
Venkatesan, Hari	13	Teaching Translation in the Age of Neural Machine Translation	2018	Conference Paper	RTL
Doherty, Stephen / Moorkens, Joss / Gaspari, Federico / Castilho, Sheila	14	On Education and Training in Translation Quality Assessment	2018	Book Chapter	RTL, CSL

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Blagodarna, Olena		Enhancement of post-editing performance: introducing machine translation post-editing in translator training	2018	PhD Dissertation	Rejected
Kim, Dohun / Koh, Taejin		Tandem Translation Classroom: A Case Study	2018	Journal Article	Rejected
Hu, Wan		Revisiting Translation Quality Assurance: A Comparative Analysis of Evaluation Principles between Student Translators and the Professional Trans-Editor	2018	Journal Article	Rejected
Çetiner, Caner		Analyzing the Attitudes of Translation Students towards CAT (Computer-Aided Translation) Tools	2018	Journal Article	TPM, CSL
Esqueda, Marileide Dias		ENSINO DE TRADUÇÃO: CULTURAS PEDAGÓGICAS	2018	Journal Article	Rejected
Sales, Dora / Pinto, María / Fernández-Ramos, Andrés		Undressing information behaviour in the field of translation: A case study with Translation trainees	2018	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Castilho, Sheila / Doherty, Stephen / Gaspari, Federico / Moorkens, Joss	15	Approaches to Human and Machine Translation Quality Assessment	2018	Book Chapter	RTL
Wu, Xiaoyan / Zhang, Shuili		Evaluation of Computer Assisted Translation and its Teaching Application in English Translation	2018	Journal Article	Rejected
Moorkens, Joss	16	What to Expect from Neural Machine Translation: A Practical In-Class Translation Evaluation Exercise	2018	Journal Article	RTL, TPM
Rodríguez-Castro, Mónica		An Integrated Curricular Design for Computer-Assisted Translation Tools: Developing Technical Expertise	2018	Journal Article	Rejected
Stasimioti, Maria / Sosoni, Vilelmini		Undergraduate Translation Students' Performance and Attitude vis-à- vis Machine Translation and Post-editing: Does Training Play a Role?	2019	Conference Paper	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Bowker, Lynne / CIro, Jairo Buitrago		Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community	2019	Book	Rejected
Dede, Volkan		Does a formal post-editing training affect the performance of novice post-editors? An experimental study	2019	Journal Article	Reject
Hatim, B. / Munday, Jeremy		Translation: an advanced resource book for students	2019	Book	Rejected
Portilho, Talita		Avaliação de tradução nos contextos profissional e pedagógico: proposta de unidade didática para revisão e avaliação por pares	2018	Master's Thesis	Rejected
Vieira, Pedro Miguel Sapateiro		Revisão e qualidade na tradução: reflexões teóricas e práticas	2019	Master's Thesis	Rejected
Cid-Leal, Pilar / Espín-García, María-Carmen / Presas, Marisa	17	Traducción automática y posedición: perfiles y competencias en los programas de formación de traductores	2019	Journal Article	RTL, CSL

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Zhang, Hong		Formación en posedición de traducción automática para estudiantes de lenguas extranjeras	2019	PhD Dissertation	Rejected
Núñez, Kenneth Jordan	18	ANÁLISIS DE LA PERCEPCIÓN, LA UTILIDAD Y LA CALIDAD DE LOS SISTEMAS DE TA POR PARTE DEL TRADUCTOR EN FORMACIÓN	2019	Journal Article	TPM
Lara, Cristina Plaza		Análisis DAFO sobre la inclusión de la traducción automática y la posedición en los másteres de la red EMT SWOT Analysis of the Inclusion of Machine Translation and Post- Editing in the Master's Degrees Offered in the EMT Network	2019	Journal Article	Rejected
McAndrew, Ewan / Campbell, Lorna		Building Bridges Not Walls – Wikipedia in Translation Studies	2019	Journal Article	Rejected
Chien, Chi-ying		Teaching Translation with the Notion of Discourse Analysis under Different Communication Patterns	2019	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Koponen, Maarit / Salmi, Leena / Nikulin, Markku		A Product and Process Analysis of Post-Editor Corrections on Neural, Statistical and Rule-Based Machine Translation Output	2019	Journal Article	Rejected
Yamada, Masaru		The Impact of Google Neural Machine Translation on Post-Editing by Student Translators	2019	Journal Article	Rejected
Jia, Yanfang / Carl, Michael / Wang, Xiangling		How Does the Post-Editing of Neural Machine Translation Compare with From-Scratch Translation? A Product and Process Study	2019	Journal Article	Rejected
Guerberof Arenas, Ana / Moorkens, Joss		Machine Translation and Post-Editing Training as Part of a Master's Programme	2019	Journal Article	Rejected
Öner Bulut, Senem	19	Integrating Machine Translation into Translator Training: Towards 'Human Translator Competence'?	2019	Journal Article	RTL, TPM
Plaza Lara, Cristina		Análisis DAFO sobre la inclusión de la traducción automática y la posedición en los másteres de la red EMT	2019	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Rothwell, Andrew / Svoboda, Tomáš	20	Tracking Translator Training in Tools and Technologies: Findings of the EMT Survey 2017	2019	Journal Article	RTL, CSL, MOD
Nitzke, Jean / Tardel, Anke / Hansen-Schirra, Silvia		Training the Modern Translator: The Acquisition of Digital Competencies through Blended Learning	2019	Journal Article	Rejected
Rodríguez de Céspedes, Begoña	21	Translator Education at a Crossroads: The Impact of Automation	2019	Journal Article	RTL, CSL, MOD
Yang, Yanxia / Wang, Xiangling		Modeling the intention to use machine translation for student translators: An extension of Technology Acceptance Model	2019	Journal Article	Rejected
Nitzke, Jean / Tardel, Anke / Hansen-Schirra, Silvia		Training the modern translator – the acquisition of digital competencies through blended learning	2019	Journal Article	Rejected
Barros, Elsa Huertas / Vine, Juliet		Training the trainers in embedding assessment literacy into module design: a case study of a collaborative transcreation project	2019	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Wang, Feng (Robin) / Humblé, Philippe		Corpus-Informed Translation Studies– Looking Backward and Forward: A Chinese Perspective	2019	Journal Article	Rejected
Bogush, Alla M. / Korolova, Tetiana M. / Popova, Oleksandra V	22	TEACHING MACHINE TRANSLATION TO THE STUDENTS MAJORING IN THE HUMANITIES	2019	Journal Article	TPM
Udina, Natalia		USING POST-EDITING IN TRANSLATION AND LSP COURSES	2019	Conference Paper	Rejected
Turkmen, Burcu / Can, Muhammed Zahit		The Opportunities and the Limitations of using the Independent Post-Editor Technology in Translation Education	2019	Journal Article	Rejected
Gomaa, Yasser A. / AbuRaya, Rania / Omar, Abdulfattah		The Effects of Information Technology and E-Learning Systems on Translation Pedagogy and Productivity of EFL Learners	2019	Conference Paper	Rejected
Bowker, Lynne		Machine translation literacy: Academic libraries' role	2019	Poster	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Cid, Clara Ginovart / Colominas, Carme / Oliver, Antoni		Language industry views on the profile of the post-editor	2020	Journal Article	Rejected
Ginovart, Clara / Oliver, Dr Anton		The post-editor's skill set according to industry, trainers and linguists	2020	Book Chapter	Rejected
Mitchel-Schuitevoerder, Rosemary	23	A Project-Based Approach to Translation Technology	2020	Book	RE, TPM
Çetiner, Caner		The effect of post-editing on the attitudes and performance of translation students	2020	Journal Article	Rejected
Koponen, Maarit / Mossop, Brian / Robert, Isabelle / Scocchera, Giovanna (Eds.)		Translation revision and post-editing: industry practices and cognitive processes	2020	Book	Rejected
Díaz-Millón, Mar / Gutiérrez-Artacho, Juncal		NUEVOS PERFILES PROFESIONALES EN TRADUCCIÓN: COMPETENCIAS NECESARIAS PARA LA TRANSCREACIÓN	2020	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Olkhovska, Alla / Frolova, Iryna	24	Using Machine Translation Engines in the Classroom: A Survey of Translation Students' Performance	2020	Journal Article	RTL
Shraida, Jabbar Frayyeh / Khalaf, Muthana Hameed		Irrational Thoughts of the Translator's Personality in Assessing the Quality of Translation	2020	Journal Article	Rejected
K. Pokorn, Nike / Blake, Jason / Reindl, Donald / Pisanski Peterlin, Agnes		The Influence of Directionality on the Quality of Translation Output in Educational Settings	2020	Journal Article	Rejected
Nanivska, Lidiia L. / Shcherba, Oksana V. / Pasichnyk, Susanna M. / Kozyar, Mykhaylo M. / Savka, Iryna V		Future Translator Training Based on an Integrated Approach	2020	Journal Article	Rejected
Loock, Rudy	25	No More Rage against the Machine: How the Corpus-Based Identification of Machine-Translationese Can Lead to Student Empowerment	2020	Journal Article	RTL

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Yanlin, Guo	26	Prospects for the Teaching of Translation Majors in the New Era	2020	Journal Article	RTL
O'Brien, Sharon / Ehrensberger-Dow, Maureen		MT Literacy—A Cognitive View	2020	Journal Article	Rejected
Lévano Castro, Sofía		Pensamiento crítico y adquisición de la competencia estratégica en estudiantes de traducción	2020	Journal Article	Rejected
Bowker, Lynne		Machine translation literacy instruction for international business students and business English instructors	2020	Journal Article	Rejected
Bowker, Lynne		Chinese speakers' use of machine translation as an aid for scholarly writing in English: a review of the literature and a report on a pilot workshop on machine translation literacy	2020	Journal Article	Rejected
Yang, Yanxia / Wang, Xiangling		Predicting student translators' performance in machine translation post-editing: interplay of self-regulation, critical thinking, and motivation	2020	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Tang, Jun		Graduate-level career preparation for Chinese translation students: a perspective of educational ergonomics	2020	Journal Article	Rejected
Konttinen, Kalle / Veivo, Outi / Salo, Pia		Translation students' conceptions of translation workflow in a simulated translation company environment	2020	Journal Article	Rejected
Whyatt, Bogusława / Naranowicz, Marcin		A robust design of the translator's skill set: evidence for transfer of metacognitive skills to intralingual paraphrasing	2020	Journal Article	Rejected
Breyel, Christine / Grass, Thierry	27	Machine translation and "biotranslation": an unlikely marriage	2020	Journal Article	RTL
Nord, Christiane		Functional Approaches to Translation	2020	Book Chapter	Rejected
Kappus, Martin / Ehrensberger-Dow, Maureen		The Ergonomics of Translation Tools: Understanding When Less is Actually More	2020	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Stasimioti, Maria / Sosoni, Vilelmini / Chatzitheodorou, Konstantinos		Investigating post-editing effort: Does directionality play a role?	2021	Journal Article	Rejected
González Pastor, Diana	28	Introducing Machine Translation in the Translation Classroom: a Survey on Students' Attitudes and Perceptions	2021	Journal Article	RTL, CSL, TPM
Pavlović, Nataša / Antunović, Goranka		Towards acquiring post-editing abilities through research-informed practical tasks	2021	Journal Article	Rejected
Dubeda, Tomáš / Obdržálková, Vanda		Stylistic Competence in L2 Translation: Stylometry and Error Analysis	2021	Journal Article	Rejected
Algryani, Ali / Syahrin, Syerina		Utilizing Learners' Linguistic Landscape as a Pedagogical Resource in the Translation Classroom: A Case Study in the Sultanate of Oman	2021	Journal Article	Rejected
Fang, Xiaoqing / Morris, Philip		Cultivation of Student Translator Autonomy in UK Higher Education	2021	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
González Pastor, Diana / Rico, Celia		POSEDITrad: la traducción automática y la posedición para la formación de traductores e intérpretes	2021	Journal Article	Rejected
Esqueda, Marileide Dias	29	MACHINE TRANSLATION: TEACHING AND LEARNING ISSUES	2021	Journal Article	TPM
Ruano, M. Rosario Martín		Towards alternatives to mechanistic models of translation in contemporary journalism	2021	Journal Article	Rejected
Krüger, Ralph	30	Using Jupyter notebooks as didactic instruments in translation technology teaching	2021	Journal Article	RTL, RE, MOD
Marczak, Mariusz / Bondarenko, Oleksandr		Translator education in Poland and Ukraine: does the academia vs industry gap persist?	2021	Journal Article	Rejected
García-Escribano, Alejandro Bolaños / Díaz-Cintas, Jorge / Massidda, Serenella		Latest advancements in audiovisual translation education	2021	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Giczela-Pastwa, Justyna		Developing phraseological competence in L2 legal translator trainees: a proposal of a data mining technique applied in translation from an LLD into ELF	2021	Journal Article	Rejected
Plaza-Lara, Cristina		Competences of translation project managers from the academic perspective: analysis of EMT programmes	2021	Journal Article	Rejected
Torres-Simón, Ester / Pieta, Hanna / Maia, Rita Bueno / Xavier, Catarina		Indirect translation in translator training: taking stock and looking ahead	2021	Journal Article	Rejected
Wang, Yun		Intention and inversion: evaluating coursebooks on translation theories under the perspective of comparative translation studies	2021	Journal Article	Rejected
Vottonen, Erja / Kujamäki, Minna		On what grounds? Justifications of student translators for their translation solutions	2021	Journal Article	Rejected

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Kruger, Ralph	31	AN ONLINE REPOSITORY OF PYTHON RESOURCES FOR TEACHING MACHINE TRANSLATION TO TRANSLATION STUDENTS	2021	Journal Article	RTL, RE
Popova, Elena Nikolaevna / Bobrova, Svetlana Evgenievna / Maltzagov, Isa Daudovich		DEVELOPMENT OF THE COMPETENCY AND CULTURE OF USING AUTOMATIC TRANSLATION TOOLS IN FUTURE TRANSLATORS (ON THE EXAMPLE OF MASTERING THE USE OF CAT TOOLS)	2021	Journal Article	Rejected
Wu, Hongxia		Multimedia Interaction-Based Computer-Aided Translation Technology in Applied English Teaching	2021	Journal Article	Rejected
Munkova, Dasa / Munk, Michal / Benko, Lubomir / Hajek, Petr	32	The role of automated evaluation techniques in online professional translator training	2021	Journal Article	MOD, TPM

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Chernovaty, Leonid / Kovalchuk, Natalia		Post-Editing as the Means to Activate Students' Thinking and Analytical Process: Psycholinguistic Aspects	2021	Journal Article	Rejected
Phase 2 - Natural Language Processing					
AUTHOR(S)	NUMBER	TITLE	YEAR	TYPE OF DOCUMENT	CODE
Aldescu, Iuliana		Assessing the Student's Knowledge in Informatics using the METEOR metric	2014	Journal Article	Rejected
Dobre, Iuliana		A Comparison Between BLEU and METEOR MEtrics Used for Assessing Students within an Informatics Discipline Course	2015	Journal Article	Rejected
Alm, Cecilia Ovesdotter / Womack, Kathryn / Haake, Anne / Engstrom, Timoth	33	A Pedagogical Model for Computational Linguistics across Curricular Boundaries	2016	Journal Article	CSL

Table B.1 continued from previous page					
Phase 1 - Translation Studies					
Martynova, Irina / Metelkova, Lilia / Gordeeva, Natalia / Nikitinskaya, Larisa / Emelianova, Margarita / Trukova, Alena	34	THE PROGRAMS OF COMPUTATIONAL LINGUISTICS GRADUATE IN GERMAN AND US UNIVERSITIES	2018	Journal Article	CSL
Norouzi, Narges / Chaturvedi, Snigdha / Rutledge, Matthew		Lessons Learned from Teaching Machine Learning and Natural Language Processing to High School Students	2020	Journal Article	Rejected
Artemova, Ekaterina / Apishev, Murat / Sarkisyan, Veronika / Aksenov, Sergey / Kirjanov, Denis / Serikov, Oleg	35	Teaching a Massive Open Online Course on Natural Language Processing	2021	Journal Article	CSL, MOD

Appendix C

Phase 1b - Curriculum Analysis - Programmes Analysed

Table C.1: Universities and Master's programmes analysed

Region / Country	**University**	**Master's Programme**
Europe		
Germany	University of Konstanz	Master's in Speech and Language Processing
Germany	Ruprecht-Karls-University of Heidelberg	Master in Computational Linguistics
Germany	Ludwig-Maximilians-University (LMU)	M.Sc. in Computer Linguistics
Germany	Saarland University	M.Sc. in Language Science and Technology
EU / Joint	LCT Consortium	European Master in Language and Communication Technologies

Continued on next page

Region / Country	University	Master's Programme
Austria	University of Vienna / FH Campus Wien	Multilingual Technologies
Belgium	KU Leuven	Master of Advanced Studies in Linguistics – NLP Track / Master of Artificial Intelligence
Belgium	University of Antwerp	Master in Computer Science: Data Science and AI / Master in Digital Text Analysis
Czech Republic	Charles University	European Master in LCT / Master in Language Technologies and Computational Linguistics
Poland	Silesian University	Master in Computer Science
Denmark	University of Copenhagen	Master in Computer Science
Denmark	University of Southern Denmark	Master in Computer Science
Estonia	University of Tartu	Master in Computer Science
Finland	University of Helsinki	Master in Data Science
Germany	Universität Tübingen	MA in Computational Linguistics / Master in Informatics
Germany	University of Bremen	Master in Informatics
Germany	University of Düsseldorf	Master in Informatics
Greece	Athens University of Economics and Business	Master in Computer Science

Continued on next page

Region / Country	University	Master's Programme
Ireland	Dublin City University	Master in Natural Language Processing
Italy	Università di Bologna	Master in Natural Language Processing
Italy	Università di Padova	Master in Information and Data Engineering
Italy	Università di Siena	Master in Artificial Intelligence and Automation Engineering
Malta	Università ta' Malta	M.Sc. in Human Language Science and Technology
Norway	Norwegian University of Science and Technology	Master of Informatics
Poland	University of Warsaw	Master in Computer Science
Portugal	University of Lisboa	Master in Computer Networks and Engineering
Romania	Alexandru Ioan Cuza University	Master in Computational Linguistics
Slovenia	University of Ljubljana	Master in Computer and Information Science
Spain	University of the Basque Country	Master in Language Analysis and Processing
Spain	Universitat Autònoma de Barcelona	Master in Research and Innovation in Computer-Based Science and Engineering
Spain	Universidad da Coruña	Master in Artificial Intelligence

Continued on next page

Region / Country	University	Master's Programme
Sweden	Göteborg University	Master in Language Technology
Sweden	Linköping University	Master in Computer Science and Engineering / M.Sc. in Information Technology
Sweden	Lund University	Master in Machine Learning, Systems and Control
Sweden	Uppsala University	Master in Language Technology
Sweden	Växjö University	Master in Software Technology
Switzerland	University of Zurich	Master in Informatics
United Kingdom		
UK	Imperial College London	Master in Artificial Intelligence
UK	Loughborough University	Master in Artificial Intelligence
UK	Queen Mary University of London	Master in Artificial Intelligence
UK	University of Birmingham	Master in Artificial Intelligence and Machine Learning
UK	University of Brighton	Master in Computer Science
UK	University of Cambridge	MPhil in Advanced Computer Science
UK	University of Edinburgh	Master in Artificial Intelligence
UK	University of Essex	Master in Computer Science
UK	University of Leeds	Master in Computer Science
UK	University of Sheffield	Master in Computer Science
UK	University of Sussex	Master in Artificial Intelligence and Adaptive Systems

Continued on next page

Region / Country	University	Master's Programme
UK	Ulster University	Master in Computer Science
North America		
USA	Stanford University	Postgraduate NLP/ML-related programmes
USA	University of Washington	Master in Computational Linguistics
USA	Carnegie Mellon University	NLP/ML-related Master's
USA	University of Colorado	MS in Computational Linguistics, Analytics, Search and Informatics
USA	Brandeis University	MS in Computational Linguistics
USA	Georgetown University	MS in Computational Linguistics
USA	Indiana University	MA / MS in Computational Linguistics
USA	University of California, Santa Cruz	MS in Natural Language Processing
USA	University of Rochester	MS in Computational Linguistics

Appendix D

Phase 1c - Questions for Survey

1. Q1 - Name - (Open-ended)
2. Q2 - Email - (Open-ended)
3. Q3 - List of Countries - (Close-ended)
4. Q4 - What is your highest level of education? - (Close-ended)
5. Q5 - What is your field? - (Close-ended)
6. Q6 - How many hours do you spend teaching per week? Move the slider according to the amount of hours. - (Close ended)
7. Q7 - What are your other main work activities? - (Close-ended)
8. Q8 - In your current teaching role, how much influence do you have over the curriculum, including changes to the syllabus and teaching methods? - (Close-ended)
9. Q9 - Please, add any further comments or explanations for your previous answer here. - (Open-ended)
10. Q10 - What type of students do you work with, mostly? - (Close-ended)
11. Q11 - At what academic levels do you currently teach? Please select all that apply. - (Close-ended)

12. Q12 - What modality/modalities do you deliver training in? - (Close-ended)

13. Q13 - Have you taught MT quality assessment before? - (Close-ended)

14. Q14 - Please rate the significance of incorporating human evaluation into the development of MT systems. Rate on a scale of 1 to 5. - (Close-ended)

15. Q15 - Please, add any further comments or explanations for your previous answer here. - (Open-ended)

16. Q16 - In your opinion, what trends do you foresee in evaluation metrics that incorporate human judgement for MT systems? Select all that apply. - (Close-ended)

17. Q17 - In your view, what constitutes a comprehensive evaluation of an MT system? Please describe the key components or criteria that should be included. - (Open-ended)

18. Q18 - What types of MT evaluation do you teach? - (Close-ended)

19. Q19 - Please, add any further comments or explanations for your previous answer here. - (Open-ended)

20. Q20 - Do you teach evaluation for NLP tasks (e.g. summarisation, speech recognition, sentiment analysis) other than MT? - (Close-ended)

21. Q21 - Please, add any further comments or explanations for your previous answer here. - (Open-ended)

22. Q22 - How many years have you been teaching MT quality assessment? Move the slider according to the amount of years. - (Close-ended)

23. Q23 - Assess the importance of teaching students how to plan evaluations for MT systems in your academic curriculum. Please rate the importance of integrating evaluation planning as part of the academic curriculum for MT quality assessment. - (Close-ended)

24. Q24 - Please, add any further comments or explanations for your previous answer here. - (Open-ended)

25. Q25 - Assess the importance of including Evaluation Metrics in your academic curriculum. Please evaluate the importance of integrating evaluation metrics into the academic curriculum for MT quality assessment. You will be presented with two types of evaluation metrics. Rate on a scale of 1 to 5. - (Close-ended)

26. Q26 - Please, add any further comments or explanations for your previous answer here. - (Open-ended)

27. Q27 - Beyond content (such as human evaluation metrics or automatic evaluation metrics), what other pedagogical aspects do you believe may be currently lacking in the teaching of MT quality assessment? Please select all that apply. - (Close-ended)

28. Q28 - Please, add any further comments or explanations for your previous answer here. - (Open-ended)

29. Q29 - If you were to create a Human Evaluation module in MT quality assessment addressed to NLP students at Master's level, what should be the main content? Select as many as necessary. - (Close-ended)

30. Q30 - Please, add any further comments or explanations for your previous answer. - (Open-ended)

31. Q31 - If you were to create a Human Evaluation module in MT quality assessment addressed to NLP students at Master's level, what would be the best format? -(Close-ended)

32. Q32 - Given your previous choice on the best format for a Human Evaluation module, what teaching modality would be most suitable? - (Close-ended)

33. Q33 - Please, add any further comments or explanations for your previous answers from Q31 and Q32 here. - (Open- ended)

34. Q34 - Is there anything else you would like to add? - (Open-ended)

Appendix E

Phase 3 - Syllabus Development

UNITS	LEARNING OUTCOMES	ACTIVITIES	WORK LOAD	CONTENT	ASSESSMENT
What is the title or theme of this unit?	What should the student be able to do, demonstrate, or apply by the end of the unit?	What will the student do to meet the outcomes?	Lecture and Lab Workload required.	What concepts, skills, or materials will the student be exposed to in this unit?	How will the student be assessed?

Unit 1 - Using Metrics Critically Pt.1	Develop and foster awareness of the role of human evaluation in the development of Machine Translation systems.	Task 1 - Human Evaluation - Practice of Adequacy, Fluency, Error Classification.	Lecture: 2h, Lab: 2h	Purpose of performing quality assessment. Strengths and weaknesses of human and automatic evaluation.	Reflective Report (200-500 words) on translation quality and evaluation approaches.
Unit 2 - Using Metrics Critically Pt.2	Explore the role of automatic evaluation metrics in the development of machine translation systems.	Task 2 - Measuring Human Evaluation - Inter-Annotator Agreement - Unweighted Cohen's Kappa.	Total: 4 hours	Exploration of human parity claims and limitations of Sentence-Level evaluation in MT.	Assessment 2: Reflective Report (200-500 words) on document-level evaluation metrics and their utility.

Unit 3 - Evaluation Design	Develop an understanding of key elements in the design of machine translation systems evaluation, including ethical considerations.	Task 1 - Evaluation Design: Readings and planning on content type, metrics, evaluator type, and platforms.	Total: 4 hours	Discussion on the importance of evaluation design, evaluator selection criteria, and international standards.	Assessment 3: Group-based Machine Translation Evaluation Planning and reporting.
Unit 4 – Ethics in Evaluation	Report results from evaluations of machine translation systems, addressing diverse evaluation approaches.	Task 1 - Case Studies: Identifying AI hype and evaluation flaws.	Total: 4 hours	Introduction to ethical implications in machine translation evaluation.	Assessment 4: Group-based Case Study on dismantling AI hype and flawed MT evaluations.

Appendix F

Phase 3 - Course Development

F.1 Lesson Plan 1 - Lecture

Lesson Plan 1 - Lecture

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 1: Warming Up - Introduction	The teacher will start the lesson by introducing themselves. The teacher will ask students to introduce themselves and share their expectations for the training. The teacher will also ask students if they speak languages other than English, noting potential language pairs that might be useful for the labs.	-
Topic 2: Consent Form	The teacher will inform students about the research project and invite them to participate. The teacher will explain in plain terms what participation entails and then distribute the consent form for students to read.	Attachment I - Consent Form

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 3: Questionnaire	The teacher will provide the students who accepted to participate in the research with a QR code to access a questionnaire (Attachment II). The teacher will explain to students that should they have any questions about the research or the questionnaire, they are able to ask the teacher, who is also acting as the researcher.	Attachment II - Questionnaire, Slides
Topic 4: What Do You Know About Quality Assessment?	The teacher will proceed with the lesson and open the slides. As a warm-up, the teacher will ask students what they know about quality assessment. The teacher may ask students to recall what they filled in the questionnaire to explain their perspectives.	Slides 4 to 8

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 5: Quality - Evaluation with Humans	The teacher will ask students how quality can be measured with humans. The teacher will ask about the strengths and weaknesses of evaluation with humans. For each topic that the students mention, the teacher will: 1. Write it on the board or show it on the slides. 2. Ensure the following points are addressed if students miss them: - Detailed evaluations. - Complex linguistic phenomena. - Deeper analyses of system performance. - Feedback to the MT systems. - Diagnoses. - Avoiding the Artificial Intelligence (AI) hype. If students bring up examples from automatic metrics, the teacher will highlight their weaknesses or focus on the strengths of human evaluation.	Slides 9 to 12

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 6: Purpose of Evaluating an MT System	The teacher will explain that, in addition to the strengths and weaknesses of human evaluation, the purpose of the evaluation also matters. While the teacher will discuss this in more detail in Lesson 3, they will introduce the concept in Lesson 1. The teacher will explain that the evaluator's profile should match the purpose of the evaluation, such as: - End-user (e.g., gisting or dissemination). - Post-editor (e.g., light or heavy post-editing, productivity, or satisfaction factors). - Specific applications (e.g., multilingual information retrieval). - Tuning or diagnosis for system improvement.	Slides 13 to 14
Topic 7: The AI Hype	The teacher will discuss AI hype with the students, focusing on how AI-based technologies have generated overblown expectations about the capabilities of MT systems. The teacher will explain that proper evaluation with humans can help dispel the hype, particularly claims that MT systems can replace translators. The teacher will clarify that, no matter how sophisticated AI systems become, they may never completely replace humans, especially for critical tasks.	Slides 15 to 21

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 8: Evaluation Flowchart	The teacher will show a flowchart from Doherty et al. (2018), outlining what type of evaluation should be conducted in different scenarios. The teacher will highlight the date of the flowchart, discussing how research has expanded metrics and introduced new ones since its publication. If necessary, the teacher may suggest taking a 10-minute break after this topic.	Slides 22 to 24
Topic 9: Adequacy/Fluency	The teacher will explain the use of adequacy and fluency measures. Key points to discuss: - Adequacy (also known as accuracy or fidelity) usually focuses on the source text, while fluency (also known as intelligibility) focuses on the target text. - These metrics are often measured together using Likert scales. The teacher will provide examples of Likert scales, such as: - Fluency: "1. No fluency; 2. Little fluency; 3. Near native; 4. Native." - Adequacy: "1. None of it; 2. Little of it; 3. Most of it; 4. All of it." The teacher will briefly introduce Likert scales and note that their implementation will be discussed in more detail during Lesson 2.	Slides 25 to 27

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 10: Error Classification	The teacher will explain that error classification metrics aim to identify and categorise translation errors. The teacher will mention various proposed taxonomies but emphasise that MQM-QT21, particularly as DQF-MQM, is the most widely adopted. Examples of error types, such as word order and mistranslation, will be discussed, as these may be encountered during the labs.	Slides 28 to 31
Topic 11: Acceptability	The teacher will explain that acceptability is a metric used in various fields, such as linguistics, translation, and human-computer interaction (HCI). The teacher will highlight that it is challenging to find a single definition of acceptability, as interpretations may vary across different authors and disciplines. The discussion will introduce how acceptability metrics can apply to MT evaluation.	Slides 32 to 33

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 12: Ranking	The teacher will explain that ranking can be applied from different perspectives. Key points to discuss: - For translators as expert evaluators, ranking complements other metrics when comparing different systems. - For end-users, ranking can be used in usability evaluations. The teacher will conduct a quick exercise where students rank three MT outputs in English, focusing on fluency or other aspects. The activity will help students understand the practical application of ranking in MT evaluation.	Slides 34 to 35
Topic 13: Usability	The teacher will explain how usability metrics can be applied in MT evaluation, focusing on the final readers of the translation, especially end-users. Key points to cover: - Usability includes objective measures, such as task completion time and browsing behaviour. - It also includes subjective measures, like user opinions and satisfaction. The teacher will highlight tools for monitoring user interactions with MT systems, such as screen recording, keyboard logging, and eye tracking.	Slide 36

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 14: Inter-Annotator Agreement (IAA)	The teacher will explain that Inter-Annotator Agreement (IAA) measures how consistently multiple evaluators make the same decisions for a given metric or category. Key points to discuss: - IAA reveals whether guidelines provided to evaluators are clear, whether the evaluation task is reproducible, and whether annotators understood the guidelines. - Commonly used coefficients include Cohen's Kappa and Fleiss' Kappa. The teacher will emphasise the importance of IAA in ensuring reliable evaluation results.	Slides 37 to 43
Topic 15: Closing Remarks	The teacher will conclude the session by asking students for their thoughts on the topics covered during the lesson. Key points to address: - Whether students were already familiar with any of the metrics discussed before the training. - Any questions or clarifications needed regarding the concepts introduced. The teacher will emphasise the importance of understanding and applying these evaluation metrics effectively in MT quality assessment.	Slides 44 to 45

F.2 Lesson Plan 2 - Lab

Lesson Plan 2 - Lab

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 1: Warming Up - Instructions	The teacher will ask students to open two documents. These documents will have been sent by email: i) Attachment I - Instructions - A PDF document explaining the order of activities for the lab and how the tasks should be performed. ii) Attachment II - Metrics Sheet - A Google Sheets document with different tabs. Each tab will have the name of the metrics investigated in the lab: Adequacy/Fluency, Usability, Inter-Annotator Agreement (IAA). The teacher will explain that these documents are self-contained, and should the students lose track of their progress, to refer to the instructions when necessary. There is a teacher sheet with all the information regarding the systems, responses and other information that may be relevant for the lesson. The lesson is divided into two parts, with Adequacy, Fluency, Errors, IAA being practised from the perspective of translators. Comprehensibility, Usability (a post-task questionnaire) and Ranking being practised from the perspective of an end-user of MT.	Attachment I - Instructions, Attachment II - Metrics Sheet

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 2: Practice of Metrics - Adequacy, Fluency, Errors	The teacher will ask students to perform the first task, related to Adequacy, Fluency and Errors. The students will see 18 sentences translated from Brazilian Portuguese into English. This choice was deliberately made with the assumption that students might be monolingual. Three systems were used to translate the source language, Google MT, Bing MT and eTranslation. The teacher sheet has the specific segments which were used by each MT system. The students, however, do not have that information. Students will then rate Adequacy (if possible, considering they could be monolingual, they may ignore this measure). Answering the question "How much of the meaning expressed in the source appears in the translation", on a scale of 1 to 4, 1 being None of it and 4 being All of it, students will choose accordingly. Following that, Fluency will be measured. Answering the question "How fluent is the translation", on a scale of 1 to 4, 1 being "No Fluency" and 4 being "Native", students will choose accordingly. Finally, on errors, students may attempt to find errors in MT output, choosing from "mistranslated", "untranslated", "word order", "word form" and "no error".	Task 1 tab

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 3: Practice of Metrics - Inter-Annotator Agreement (IAA)	In the following task, students will be asked to calculate their IAA. Clarify that the IAA being measured in this lesson is only the Unweighted Cohen's Kappa. In the following class other measures of IAA will be taught. The teacher will ask students to form a pair with a colleague and compare scores. Students are supposed to observe their levels of agreement for Adequacy (if performed) and Fluency. The teacher should walk by each desk consulting students on their scores.	Task 2 tab
Activity 4: Discussion	The teacher will promote a discussion based on certain aspects that might come up throughout the first part of the lesson: If monolingual, how challenging was it to evaluate without the source text? If not monolingual, how did the overall difficulty compare? Is there a correlation between Adequacy/Fluency scores and the number of errors identified? Specifically, did lower scores in Adequacy/Fluency correspond with more errors being identified, and did higher scores correspond with fewer errors? How would a deep understanding of both languages impact the evaluation? What would happen if someone without any skills attempted to assess the quality? When calculating IAA, what discrepancies (if any) were noticed in the assessments? What do you think caused them?	-

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 5: Practice of Comprehensibility	Following the previous activity, inform the students from the next task on, they will perform it from the perspective of an end-user. Explain to students that the context of these sentences is that they are taken from product review and business review websites. They will see the original review, with the translated version being made available for them, as the end-users. Students are supposed to read each translated sentence taken from a review and respond to "What type of description or judgement can you detect reading the review?", choosing among the following options: "A positive judgement", "A negative judgement", "Both", and "Neither positive nor negative".	Task 3 tab
Activity 6: Usability Post-task Questionnaire	Following the comprehensibility activity (task 4), students will be guided to respond to a post-task questionnaire on Usability. The following statements are given, and students have to agree or disagree: i) The reviews were clear and easy to understand. ii) I am satisfied with the quality of the translated reviews. iii) The translated reviews were overall comprehensible. iv) I would be able to understand similar reviews again in the future. v) The overall usability of the translated reviews was high. Students can then choose from 1. Strongly Disagree up until 5. Strongly Agree.	Task 4 tab

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 7: Practice of Ranking	The teacher will then ask students to go to tab 5, where students will measure ranking. Students will be presented with the MT output of two MT systems, still on the topic of product reviews. Students are allowed to choose between MT1, MT2 or both. Sentences were either translated by Google Translate or Amazon Translate on MT1 and Bing was used for MT2.	Task 5 tab

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 8: Discussion	The teacher will promote a discussion based on certain aspects that might come up throughout the second part of the lesson: How challenging was comprehensibility? Considering different languages were evaluated, do you think the language affected the MT? Open up a discussion on low resource languages and how that impacts the quality. What could be the potential consequences for an end-user if a translation inaccurately conveys the sentiment of a review? How could this affect an end-user's purchasing decision or perception of a business? What risks are associated with poor translation quality in this context? Regarding usability, what role does ease of understanding and use play in the perceived reliability of the information. How does the concept of usability in translation differ when considered from an end-user's perspective versus a translator's perspective? How would usability be tested for a translator? (Guide students towards other measures, such as efficiency when post-editing). How was it to rank translations without a clear criteria? Would it be better to implement criteria?	-

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 9: Reflective Report	The teacher will ask students to write a reflective report ranging from 200 to 500 words. The purpose of the report is to reflect on their experience with the different exercises performed during the lesson. The following questions will be presented on a template to guide their reflection: What were your expectations in terms of translation quality before doing this exercise? How did you adapt your evaluation approach when considering the translator's perspective versus the end-user's perspective? Did you expect the IAA results you have obtained? Comment on the differences between your scores and of your colleague's. Will this change how you use MT?	Reflective Template

F.3 Lesson Plan 3 - Lecture

Lesson Plan 3 - Lecture

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 1: Warming Up	The teacher will begin the lesson by asking students how they are doing. The teacher will ask students about the reflective assignment from the previous lab, in case any student had not finished at the lab.	-

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 2: Has Machine Translation Achieved Human Parity?	The teacher will proceed with the lesson and open the slides. As a warm-up, the teacher will ask students what they know about the concept of "human parity." The teacher will mention studies that claimed that Machine Translation (MT) has achieved human parity. The teacher will explain the claim made by Hassan et al. (2018) and how other studies attempted to analyse those claims (Toral et al., 2018; Läubli et al., 2020; Graham et al., 2020). The teacher may mention some contributions from these studies, but they will be analysed in depth in Lesson 3 for MT evaluation design.	Slides 4 to 12

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 3: On Context for Machine Translation	Because the capabilities of MT must be checked, the teacher will turn the students toward a discussion on the current paradigm of MT, which is sentence-level. The teacher will ask students about the previous lab, specifically if at any point the sentences were difficult to evaluate due to the lack of context. The teacher will show studies that have attempted to incorporate discourse into current MT systems (Wang, 2019). The teacher will explain different approaches, such as: - Systems that rely on a context span limited to sentence pairs (Tiedemann and Scherrer, 2017; Bawden et al., 2018; Müller et al., 2018). - Systems that rely on document substructures (Dobreva, Zhou, and Bawden, 2020). - Systems that attempt to go beyond the span (Junczys-Dowmunt, 2019; Voita, Sennrich, and Titov, 2019).	Slide 13

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 4: Features of Context	The teacher will explain to students that work done in Translation Studies has described aspects of context based on the study by Melby and Foster (2010): - Co-text. - Chron-text. - Rel-text. - Bi-text. - Non-text. The teacher will explain how these concepts are applied in MT development and their relevance. The teacher will briefly explain how these topics relate to MT evaluation, covering: - Intersentential context. - World knowledge and external information. - Terminology.	Slides 14 to 15
Topic 5: Efforts into Investigating Document-Level NMT	The teacher will report different surveys published in the field, demonstrating efforts made to discuss and report methodologies and model architectures. The teacher will report the contributions of: - Popescu-Belis (2019). - Kim, Tran, and Ney (2019). - Lopes et al. (2020). - Maruf, Saleh, and Haffari (2021). - Jin et al. (2023). - Castilho and Knowles (2024). The teacher will guide students through these contributions and explain their relevance to document-level NMT research.	Slides 17 to 22

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 6: Human Evaluation and Semi-Automatic Evaluation and Context	The teacher will report findings of research done in terms of human evaluation and semi-automatic evaluation pertaining to context. The teacher will report studies that investigate: - New methodologies for assessment. - Context needed for translators. - Implementation of Inter-Annotator Agreement (IAA). - Test suites. - Measures of post-editing with context in mind. When reporting test suites, the teacher will explain to students that these will be discussed in more detail in Lesson 3, focusing on MT evaluation design.	Slides 23 to 31

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 7: Automatic Evaluation Metrics and Context	The teacher will report findings of automatic evaluation metrics focusing on context. Before introducing context-specific metrics, the teacher will review sentence-based metrics, such as: - BLEU (Papineni, 2002). - METEOR (Banerjee, 2005). - chrF (Popovic, 2015). - TER (Snover, 2006). - BERTScore (Zhang, 2019). After that, the teacher will discuss the latest document-level metrics, including: - Applying sentence-level metrics to a full document through concatenation (Wong and Kit, 2012; Gong, Zhang, and Zhou, 2015; Xiong et al., 2019; Liu et al., 2020; Saunders, Stahlberg, and Byrne, 2020). - BlonDe (Jiang et al., 2022). - Building sentence representations from context and integrating them into pre-trained MT metrics, such as BERTScore and Prism (Thompson and Post, 2020). - DiscoScore (Zhao et al., 2023). - SLIDE (Raunak et al., 2023a). - Doc-COMET (Vernikos et al., 2022). The teacher will explain the strengths and weaknesses of these approaches.	Slides 32 to 41

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 8: Closing Remarks	The teacher will conclude the session by asking students for their thoughts. Key discussion points: - If students had prior knowledge of any document-level approaches to human evaluation or automatic evaluation metrics mentioned in the lesson. - Any remaining questions or clarifications needed. The teacher will thank students for their participation and provide a brief overview of the next session.	-

F.4 Lesson Plan 4 - Lab

Lesson Plan 4 - Lab

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 1: Warming Up - Instructions	The teacher will ask students to open one document and several links throughout the lesson. The document will be available in a Google Drive folder which will be shared before the class: i) Attachment I - Instructions: A PDF document explaining the order of activities for the lab and how the tasks should be performed. ii) Attachment II - Qualtrics: A Qualtrics survey link. It contains several sentences and texts that will be evaluated. **Link:** Document-Level Evaluation iii) Attachment III - Jupyter Notebook - BlonDe: Code to run the document-level metric BlonDe. **Jupyter Notebook:** MTQATraining-BlonDe.ipynb iv) Attachment IV - Jupyter Notebook - Inter-annotator agreement (IAA) measures: A Jupyter Notebook on Google Colab that contains code to run different IAA measures, such as Cohen's Kappa, Fleiss' Kappa, Krippendorff's Alpha and Inter-Rater Agreement. **Jupyter Notebook:** MTQATrainingIAA.ipynb The teacher will explain that these documents are self-contained, and should the students lose track of their progress, they should refer to the instructions when necessary. The texts contained in the Qualtrics link to be evaluated are six different texts that have been decomposed into sentences (for sentence-level evaluation), pieces of text (for document span evaluation), and whole texts (for full document evaluation).	Attachment I - Instructions, Attachment II - Qualtrics, Attachment III - Jupyter Notebook

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 2: Practice of Metrics - Sentence	The teacher, following the Qualtrics survey, will focus on the first part, with the evaluation of 14 questions focusing on sentences. The sentences are extracted from six different texts: Text 1 (literary), Text 2, 3, 4, 5 (general), and Text 6 (scientific). The teacher has adjusted the evaluation of Adequacy/Fluency to encompass a scale of 1 to 5. A middle point has been established, to add the option of students to be "neutral." The teacher should hold a discussion with the students after this step to check in. The discussion may be able to gather thoughts and attitudes from students. Make sure to include in the discussion: - The Likert Scale has a middle option, allowing the evaluator to choose a "neutral" option. - The different platforms being used (Google Sheets vs Qualtrics comparison). - Question 14 has a grammatical issue that can be solved within the sentence. Therefore, it does not need further context to fix it.	Attachment II - Qualtrics

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 3: Practice of Metrics - Context Span	In the second part of the evaluation on Qualtrics, students will evaluate sentences with additional context span, across six texts. The teacher has adjusted the evaluation of Adequacy/Fluency to encompass a scale of 1 to 5. **For Text 1, a literary text, specifically a short story:** - One question with one preceding and one succeeding sentence. - One question with two preceding and two succeeding sentences. - One question with three preceding and three succeeding sentences. **For Text 2, a general text, specifically a social media post:** - One question with one succeeding sentence. **For Text 3, a general text, a product review:** - One question with three succeeding sentences. - One question with two succeeding sentences. **For Text 4, a general text, a product review:** - One question with two preceding sentences. **For Text 5, a general text, a recipe:** - One question with one succeeding sentence. - One question with one preceding and one succeeding sentence. - One question with one preceding sentence. **For Text 6, a scientific text, a thesis abstract:** - One question with two succeeding sentences. - Three questions with one preceding sentence.	Attachment II - Qualtrics
	The teacher should hold a discussion with students to check in and gather their thoughts, addressing topics like	

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 4: Practice of Metrics - Document Level	Following the second part of the evaluation on Qualtrics, students will evaluate sentences with additional context span, of six texts. Additionally, students will be asked to rate the entire document as well. The teacher has adjusted the evaluation of Adequacy and Fluency to a direct assessment ranging from 1 to 100. Each scale has different settings, in order to make students test different specifications: **Text 1:** One increment (100), two labels ("None" and "All of it"), no custom position, no value shown to evaluators. **Text 2:** Four increments (25, 50, 75, 100), two labels ("None" and "All of it"), no custom position, values shown to evaluators.	Attachment II - Qualtrics
Activity 4 (cont.)	**Text 3:** Ten increments (10, 20, 30, ..., 100), two labels ("None" and "All of it"), no custom position, no value shown to evaluators. **Text 4:** One increment (100), five labels ("None", "Little of it", "Some of it", "Most of it", "All of it"), no custom position, no value shown.	Attachment II - Qualtrics

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 4 (cont.)	**Text 5:** Four increments (25, 50, 75, 100), two labels ("None" and "All of it"), custom position at 50, values shown. **Text 6:** Four increments, two labels ("None" and "All of it"), custom position at 50, no value shown, scale snaps to increments. **Discussion points:** The teacher should hold a discussion with the students after this step to check in. The discussion may be able to gather thoughts and attitudes from students. Make sure to include: - Ask students to discuss the differences between the settings of each text. - Point out to students the strengths and weaknesses of each feature (e.g., how snapping values in evaluation towards the increments might make it easy to calculate results but might not represent accurately how evaluators feel). - Ask students about possible difficulties when using scales, e.g., if the evaluator decides to use their phone, causing bad User Experience (UX). Comment about guidelines to prevent these scenarios. - Ask students to discuss the experience of evaluating the whole document and evaluating the document itself (comment about evaluation fatigue or adding complexity with the whole document being available).	Attachment II - Qualtrics

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 5: Correlating Human Evaluation with Sentence-Level Automatic Metrics - MATEO	The teacher will provide students with the link to a Google Drive folder containing files (source, target, and reference) for the six evaluations. These files are meant to be used with MATEO (Vanroy et al., 2023), either by using the platform or the code provided. Students are instructed to use the platform by uploading the provided files or running the code through the example implementations provided by MATEO. **Link:** MATEO Website Students will then explore sentence-level correlations between their human evaluations (Adequacy, Fluency) and metrics such as BLEU, METEOR, and TER. The teacher will discuss results from a conceptual perspective, pointing out the strengths and weaknesses of each automatic metric in reflecting human judgements.	MATEO Website

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 6: Correlating Human Evaluation with Document-Level Automatic Metrics - BlonDe	The teacher will provide students with the link to a Jupyter Notebook hosted on Google Colab. This notebook contains all the steps necessary to calculate the document-level BlonDe metric (Jiang et al., 2022). Students will follow the instructions in the notebook to evaluate their document-level data. The analysis will involve examining how the document-level metric aligns with their human evaluations and understanding its advantages and limitations. **Jupyter Notebook:** MTQATraining-BlonDe.ipynb The teacher will then facilitate a discussion to explore the differences between sentence-level metrics (discussed in Activity 5) and document-level metrics like BlonDe. The discussion will focus on the contextual nuances captured by BlonDe, its alignment with human evaluation, and scenarios where it may fail.	Jupyter Notebook

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 7: Inter-Annotator Agreement	The teacher will provide students with a Jupyter Notebook hosted on Google Colab. This notebook contains step-by-step instructions to calculate inter-annotator agreement (IAA) using various metrics such as Cohen's Kappa, Fleiss' Kappa, Krippendorff's Alpha, and Inter-Rater Agreement (IRR). The teacher will guide students through examples included in the notebook: - Weighted Cohen's Kappa: 2 annotators, scale 1-4, segments 1-10. - Fleiss' Kappa: 3 annotators, scale 1-5, segments 1-10. - Krippendorff's Alpha: 3 annotators, scale 1-100, segments 1-10. **Jupyter Notebook:** MTQATrainingIAA.ipynb After testing the examples provided in the notebook, students will apply one of the IAA metrics to their own evaluations from this lesson. They will calculate agreement scores for their peers' annotations and discuss their findings in class. The teacher will facilitate a discussion on: - The practical importance of inter-annotator agreement in evaluation tasks. - Strengths and weaknesses of the different IAA metrics (e.g., when weighted measures are more appropriate). - Common challenges faced during the calculation of IAA scores, such as discrepancies in annotator scales or inconsistencies in judgement.	Jupyter Notebook

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 8: Closing Remarks	The teacher will conclude the session by engaging students in a reflective discussion about the concepts and tools covered in the lesson. The teacher will ask the following questions to encourage participation: - What were the most useful insights gained from evaluating translation quality at both sentence and document levels? - How did the automatic metrics correlate with human evaluations, and what surprised you about their alignment or lack thereof? - Were there any challenges or limitations encountered when using the automatic metrics? The teacher will also address any lingering questions or concerns about the methodologies discussed during the session. Finally, students will be encouraged to consider how the tools and metrics explored in this session could be applied to their own translation evaluation workflows or research contexts.	-

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 9: Reflective Report	The teacher will ask students to write a reflective report ranging from 200 to 500 words. The purpose of the report is to reflect on their experience with the different exercises conducted during the lesson. The report will include the following guiding questions to help structure the students' reflections: - When performing document-level evaluation, what type of evaluation was more useful to you: document-span or document-level? - When performing evaluation, you had different scales (1-4, 1-5, 1-100) and different settings. Amongst these, what scale/setting could be useful to you, and what could make things more difficult? (You can think from the perspective of the evaluator or from analysing the data.)	Reflective Template

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Activity 9 (cont.)	- When interpreting automatic evaluation metrics, which type was more useful for correlating with human evaluation: sentence-level or document-level metrics? - Amongst the automatic evaluation metrics, which seemed to be more useful to you: sentence-level metrics or the document-level metric BlonDe? - How did you feel about the different types of measures for Inter-Annotator Agreement? - Will you implement any of these approaches/metrics into your practice of machine translation (MT) evaluation? Students will submit their reflective reports by the next class. The teacher will review these reflections and incorporate them into subsequent sessions for improved lesson design and tailored learning outcomes.	Reflective Template

F.5 Lesson Plan 5 - Lecture

Lesson Plan 5 - Lecture

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 1: Warming Up	The teacher will begin the lesson by asking students how they are doing. The teacher will ask students about the reflective assignment from the previous lab, in case any student had not finished at the lab.	-

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 2: Why Evaluation Design Matters	The teacher will proceed with the lesson and open the slides. The teacher will recap very briefly what was discussed in the previous lesson. The teacher will bring up again a study mentioned in the previous class, Hassan et al. (2018). The teacher will talk to students about the studies performed by Toral et al. (2018), Läubli et al. (2020), and Graham et al. (2020), which were mentioned briefly in class. The teacher will go in depth into their analyses and recommendations for MT evaluation to bring the main point: transparent and replicable quality assurance is important. The teacher will outline to students that the following aspects should be considered: - Purpose of MT system. - Evaluators. - Guidelines. - Evaluation Platforms. - Likert Scales. - Test Suites. - Risk Assessment.	Slides 4 to 12

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 3: Who Should Evaluate?	Following the recommendation of Läubli et al. (2020), the teacher will mention that professional translators should be employed for the evaluation. The teacher will explain the difficult aspects of recruitment as outlined by Doherty (2017). The teacher will explain the importance of the International Standards Organisation (ISO) and ASTM International, particularly some standards for the professionals involved in the process. The teacher will mention: - ISO 17100 - Translation services - Requirements for translation services. - ASTM F2475-23 for low-resource languages, which is relevant depending on the type of language involved. The teacher will discuss with students that end users should also be considered for evaluations, as they might be using the MT systems being developed.	Slides 13 to 22

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 4: What Are Guidelines For?	The teacher will explain what guidelines are used for. Based on Doherty (2017), the teacher will tell students that in the past it was difficult to provide detailed information on guidelines and operational definitions, but there has been a rising interest and effort into making guidelines available. The teacher will explain that guidelines can affect the validity, reliability, and efficiency of the translation. The teacher will report Knowles and Lo (2024) to state that in annotation, the behaviour of the annotator may lead to differences, such as when a task description (or guideline) does not offer enough context. The teacher will mention that evaluators should be able to ask questions about the criteria provided, time, and other requirements.	Slides 23 to 27

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 5: Where Should I Evaluate?	The teacher will explain that presentation and context are critical aspects, as a combination of user interface (UI) and a portion or the full document are relevant for an evaluation. The teacher will remind students that in the previous lesson, they used a specific interface and tested different configurations of evaluation, such as sentence-based and document-based. The teacher will list different examples from the literature where evaluation (un)friendliness might have impacted evaluations: - In crowdsourcing evaluation using Mechanical Turk, Heer and Bostock (2010) show how display configurations and even variations in luminance can impact evaluations, albeit not necessarily specific to MT. - Brivas-Iglesias and O'Brien (2023, p. 8) investigate how users feel when using MT before, during, and after, particularly with the aim of "allowing developers to include authentic human feedback in the design process in order to offer personalised tools according to the type of user." The teacher will then show the user interface of different platforms, including those used in the training before (Google Sheets, Qualtrics) and other platforms (Google Forms, Amazon Turk, Appraise, Kantan, MutNMT, MATEO).	Slides 28 to 32

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 6: What Scale Should I Use?	The teacher will remind students that evaluation typically uses Likert scales at different levels. Building on interfaces affecting evaluations, the teacher will report Knowles and Lo (2024) to explain that features of interfaces can influence the behaviour of annotators. The teacher will explain how Direct Assessment (DA) (Graham et al., 2013) has been used in WMT. The teacher will emphasise that the decision of scales must be made alongside the UI. Knowles and Lo (2024) will be used to explain calibration, particularly how normalisation of scores can be useful. The teacher will explain the distribution of scores evaluators might provide, such as: - Wide distributions (scoring anywhere from 0 to 100). - Narrow distributions (scoring mostly between certain points, e.g., 50 and 70). - Discretisation (only a few specific values are used, such as an evaluator only using 0, 50, 100). The teacher will recommend Clark et al. (2021)'s findings that can be applied to MT, emphasising: - Instructions. - Examples. - Comparison. The teacher will include Knowles and Lo (2024)'s assertion that annotators should receive instructions on how to use the scales, particularly regarding the slider, tick marks, and score space.	Slides 33 to 40

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 7: Why Should I Use Test Suites?	The teacher will report the discussion from Castilho and Knowles (2024) to discuss test suites, particularly to demonstrate how they are important for measuring how the model translates specific phenomena. While Castilho and Knowles (2024) emphasise test suites for phenomena at the document level, the teacher will mention that test suites can also be crafted for specific issues. The teacher will explain using the DELA test suite as an example and may also show examples of other test suites, including: - Rios et al. (2017). - Bawden et al. (2018). - Guillou et al. (2018). - Saunders et al. (2020). - Castilho et al. (2021). The teacher will highlight the possibility of generating variants of existing test sets based on Post and Junczys-Dowmunt's (2023) study.	Slides 41 to 48

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 8: Why Do I Need to Perform a Risk Assessment?	Following Canfora and Ottmann (2020), the teacher will inform the students about the risks that Neural MT (NMT) can produce, such as: - Generating translation errors (Koehn and Knowles, 2017). - Lack of easy solutions for liability and the use of NMT (Moorkens and Lewis, 2020). - Personal or confidential data used in free online NMT engines may be breached (Slator, 2017). The teacher will emphasise the importance of risk management, highlighting that only translators can effectively spot risks that NMT errors produce. The teacher will present to students how a combination of test suites with error detection can allow developers to notice the types of issues systems may produce.	Slides 49 to 55
Topic 9: Closing Remarks	The teacher will conclude the session. The teacher will ask students about their thoughts and if they had noticed any particular aspects from the previous two lessons that they would consider for their planning. The teacher will ask students if they have any other questions.	-
-		

F.5.1 Mid-test - Classroom quiz

Question 1 - What is the purpose of human evaluation?

Option (1) - To provide subjective opinions without structured guidelines

Option (2) - To eliminate the need of automatic evaluation, which is unreliable

Option (3) - To provide subjective opinions with structured guidelines

Option (4) - To provide insights on complex linguistic phenomena and complement automatic evaluation? (Correct)

Option (5) - To provide insights on comp´lex linguistic phenomena and eliminate automatic evaluation, which is unreliable.

Question 2 - What do reference based automatic evaluation metrics measure?

Option (1) The ability to detect tone and emotion in translations

Option (2) - The ability to detect tone, emotion, quality in translations between a machine-generated translation and a reference translation (also known as silver standard)

Option (3) - The ability to detect quality in translations between a machine-generated translation

Option (4) - The ability to measure similarity between machine-generated translations and reference translations (Correct)

Option (5) - The ability to measure similarity of machine generated translations without a reference translation

Question 3 - What do reference free automatic evaluation metrics measure?

Option (1) - The accuracy of a translation in conveying the main ideas, derived from general language models trained without specific reference texts

Option (2) - The consistency of style across multiple documents within the same domain

Option (3) - The ability of the translation to match the tone and intent of the original text as closely as possible

Option (4) - Different aspects of translation quality, depending on the features their model was trained to assess (Correct)

Option (5) - The similarity between the machine-generated translation and a translation generated by a large language model.

Question 4 - What are the advantages of performing human evaluation at the document level (or with context-span) instead of at sentence level?

Option (1) - It allows evaluators to consider broader context, but also makes things more ambiguous

Option (2) - It helps evaluators ignore minor errors that may disrupt the translation

Option (3) - It allows the measurement of adequacy, fluency and comprehensibility

Option (4) - It allows evaluators to consider broader context, which can improve assessments by eliminating ambiguity (Correct)

Option (5) - It allows evaluators to consider broader context, which can make assessments quicker

Question 5 - What are the weaknesses of sentence-level evaluation?

Option (1) - It does not allow you to do group evaluations

Option (2) - It makes it easier to overlook spelling and punctuation errors

Option (3) - It reduces the evaluator's ability to focus on meaning and focuses too much on style

Option (4) - It limits the speed and efficiency of the evaluation process

Option (5) - It ignores context across sentences, making the assessment possibly more ambiguous (Correct)

Question 6 - Select the option whcih only has human evaluation measure

Option (1) Adequacy, Fluency, BLEU, Comprehensibility, Cohen's Kappa

Option (2) Fluency, Usability, COMET, chrF

Option (3) BLEU, Ranking, Adequacy, Fluency

Option (4) Adequacy, Fluency, Ranking, Usability (Correct)

Option (5) Adequacy, Fluency, Ranking, USability, Inter-Annotator Agreement

Question 7 - What does the concept of perishability refer to?

Option (1) - Texts are meant for immediate consumption, with no purpose af-

terward

Option (2) - Texts meant to be updated regularly

Option (3) - Texts that need to be reviewed quickly to maintain accuracy

Option (4) - The action of editing machine translated texts

Option (5) - The degree to which a text is intended for immediate consumption versus having long-term value (Correct)

Question 8 - Rate your confidence level regarding MT quality assessment

Scale 1 "I feel confident using human evaluation metrics" - Scale from 1 (Strongly disagree) to 5 (Strongly agree)

Scale 2 "I feel confident using automatic evaluation metrics" - Scale from 1 (Strongly disagree) to 5 (Strongly agree)

Question 9 - What are the strengths of document-level evaluation, in your opinion? (Word cloud)

Question 10 - What challenges might arise when performing document-level evaluation, in your opinion? (Word cloud)

F.6 Lesson Plan 6 - Lab

Lesson Plan 6 - Lab

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 1: Warming Up - Instructions	The teacher will ask students to make groups. Organise students to be in at least two groups. Students will open some documents or work with physical sheets, including: - Attachment I: MT Design Sheet - Attachment II: Instruction for Students - Lesson 3 - Attachment III: MT Design Brief 1 or 2 The teacher will check if any students need assistance completing the reflective assignment from the previous lab.	Attachments I, II, III
Topic 2: Overall Assignment, Content, and Perishability	Proceeding with the lesson, the teacher will give students a brief on a MT evaluation. Based on this brief, students will plan a MT assessment using a MT design sheet. The teacher will explain each task of the instructions sheet as students work in groups. At the end of the lesson, students need to present their MT assessment design and hand in the sheets to the teacher. The teacher will remind students that this is not an exhaustive list and that other contextual factors may influence their choices. The teacher will prompt students to consider the content type and its perishability. The teacher will discuss how content can fall under different domains, such as legal documentation, creative marketing materials, or general news articles. Referring to Doherty et al., (2018), the teacher will illustrate how some content is intended for long-term use while other content may only be relevant for a short period.	-

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 3: Risk Assessment Measures	Moving forward, the teacher will guide students in assessing the risks associated with using their MT system for the chosen content type. They will explore questions such as whether the content poses any risks when used by end-users in a public setting, if there are instructions or warnings needed for subject matter experts, and the extent to which the content will be modified post-translation. Based on this risk assessment, the teacher will help students decide which human evaluation metrics to use or recommend. They will discuss involving professional translators to analyse complex linguistic phenomena, testing the system with appropriate end-users (such as physicians using MT to read scientific articles), and designing or utilizing specific test suites to ensure the system can handle critical content issues.	Attachments I, II, III
Topic 4: Test Sets	The teacher will then assist students in choosing a test set. With the domain and content selected, students will consider setting aside a portion of their corpus—for example, 10% of 1 million aligned sentence pairs—or selecting another corpus with similar content to verify the system's performance on unseen data. The teacher will explain the importance of comparing the MT system's translations with the test set to evaluate performance and how combining and correlating automatic evaluation metrics with human evaluation metrics provides a comprehensive assessment.	Attachments I, II, III

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 5: Designing Test Suites	The teacher will discuss with students whether they will use a test suite in addition to a test set. They will explore the benefits of verifying specific linguistic phenomena relevant to their content type. For instance, if building a system for medical content, students might test for terminological accuracy. If they are developing a system for gist translations but are concerned about consistency in tense and pronouns, they might test against crafted examples. Based on the hypothetical results from using a test suite, the teacher will guide students in deciding whether it's safer and more effective to fine-tune their model or train the system with more targeted data, helping them identify weaknesses in their MT system.	Attachments I, II, III
Topic 6: Choosing the right MT system	The teacher will advise students to select other MT systems capable of translating similar content types for comparison purposes. Comparing their system against others can be useful in the evaluation, as they want to measure their system's performance relative to others in executing the task for which they were commissioned.	Attachments I, II, III

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 7: Choosing Metrics	Next, the teacher will discuss the selection of appropriate metrics for quality evaluation. Emphasizing that depending on the content type and MT system, students will need to use a combination of different metrics, the teacher will guide them through the options. For established MT systems in areas like meteorology, students might emphasize automatic metrics for routine checks. If implementing the system in a new content area, they might place greater emphasis on human evaluation metrics combined with routine automatic metric checks. For developing new MT systems, the teacher will stress the importance of using a comprehensive evaluation combining human and automatic evaluation metrics to ensure the system can handle the content type effectively.	-
Topic 8: Human Evaluation Metrics	The teacher will then encourage students to consider human evaluation metrics, such as Adequacy/Fluency, Error Annotation, Readability, Comprehensibility, Usability, Acceptability, and Ranking. They will discuss how these metrics can be evaluated at the sentence level, with specific context spans, or at the document level. The teacher will remind students that they do not need to make a final decision now but should start thinking about it, as it will be impacted by later choices in their evaluation plan.	

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 9: Automatic Evaluation Metrics	Moving on to automatic evaluation metrics, the teacher will prompt students to consider different types, including reference-based metrics like BLEU, chrF, and TER; referenceless metrics like BERTScore, BLEURT, and COMET; and document-level approaches like BlonDe, DOC-COMET, DiscoScore, and SLIDE.	-
Topic 10: Evaluators	The teacher will then guide students in choosing the evaluators for their human evaluation. Emphasizing that the type of human evaluation metric must match the type of evaluator with the purpose of the evaluation, the teacher will discuss options such as expert evaluators (translators), end-users (who could be experts or non-experts), and crowdsourcing participants. They will highlight that employing translators is generally recommended, especially when developing a new MT system, if the MT system is intended for translators as users, or when performing risk assessments. The teacher will also discuss how the choice of evaluator affects the claims students can make about their system, noting that this will be further explored in the next lesson on ethics.	-

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 11: Evaluation platforms	**(5 minutes)** When it comes to choosing an evaluation platform, the teacher will explain that this decision depends entirely on the evaluation metrics and evaluators students plan to use. They will explore options such as cloud-based survey tools (Google Forms, SurveyMonkey, Qualtrics), specialized evaluation platforms (Appraise, KantanAI), and crowdsourcing platforms (Amazon Mechanical Turk) for human evaluation. For automatic evaluation, they will consider tools like MATEO and MutNMT. The teacher will caution students that while platforms and formats can be adapted, the capabilities of each platform may limit or enhance the evaluation's potential, and changes may affect the robustness of their evaluation.	-
Topic 12: Likert Scales	Discussing the use of Likert scales, the teacher will prompt students to contemplate whether they are suitable for their evaluation needs, taking into account the platform they're using and the nature of their content. They will consider scenarios where simple yes/no responses might be appropriate versus when Likert scales or Direct Assessment (DA) scales are more suitable. The teacher will guide students through decisions about including a middle point on Likert scales, showing increments or labels on DA scales, adding custom positions, deciding whether to show values to evaluators, and whether to use scale snapping. The teacher will emphasise that all these scale settings can impact the evaluation and should align with their evaluation goals.	-

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 13: Guidelines	The teacher will then explain the importance of writing clear and effective guidelines for evaluators. They will describe how guidelines serve as a bridge between the evaluation objectives and the evaluators' understanding, helping to reduce subjectivity and promote consistency. The teacher will advise students to aim for language that is clear and straightforward, outline the evaluation process step by step without overwhelming details, and offer tutorials on navigating the platform, including any specific functionalities. They will encourage students to anticipate areas where evaluators might find ambiguities and address them proactively in the guidelines. The teacher will also stress the value of conducting a pilot test before launching the full evaluation to reveal any unclear areas in the guidelines or process, allowing adjustments to be made beforehand.	-

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 14: Statistical tests	**(5 minutes)** Finally, the teacher will discuss how the evaluation will be analyzed and what statistical tests will be employed. They will explain that students will need to systematically analyze their results using statistical methods, which can be categorised into: i) Inter-annotator agreement measures such as Unweighted and Weighted Cohen's Kappa, Fleiss' Kappa, Krippendorff's Alpha, and Kendall's Tau. ii) Statistical tests for correlation such as Pearson Correlation Coefficient and Spearman's Rank Correlation Coefficient. iii) Methods for analysing statistical significance such as checking p-values, comparing means between groups, and using tests like Paired Samples t-tests, One-way ANOVA, Mann-Whitney U Test, Wilcoxon Signed-Rank Test, and Kruskal-Wallis H Test.	-
Topic 15: Presentation and Closing Remarks	**(20 minutes)** Closing the lesson, the teacher will ask students to present a summarised version of their MT evaluation plan. The teacher should take notes while they are presenting in case students might not have added in their evaluation sheet what they presented, for documentation purposes. The lesson will conclude with the teacher inviting students to share any thoughts or questions they have.	-

F.7 Lesson Plan 7 - Lecture

Lesson Plan 7 - Lecture

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 1: Warming Up	The teacher will begin the lesson by asking students how they are doing. The teacher will ask students about the reflective assignment from the previous lab, in case any student had not finished during the lab.	-
Topic 2: MT and Hype	The teacher will proceed with the lesson and open the slides. The teacher will recap very briefly what was discussed in the previous lesson. The teacher will report Moorkens, Way, and Lankford (2024) to explain that MT is far from a 'solved problem' in the technological landscape. The teacher should report a study from Terribile (2024, p. 231) that explains that over 40% of all edits in European languages required major edits. Mention as well the lack of resources in terms of data for low-resource languages. This should lead to pointing out that AI capabilities must not be overstated, in order to move towards explainable AI (xAI).	Slides 4 to 7

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 3: Transparency and Explainable AI (xAI)	The teacher will introduce the concept of transparency in AI and its importance in machine translation (MT). Key points to cover include: - The definition and goals of Explainable AI (xAI). - Why transparency is critical for building trust in MT systems. - Examples of explainable methods in MT, such as highlighting problematic translations or providing context for errors. The teacher will present slides showcasing real-world applications of xAI in MT, including visual aids to demonstrate how explainable systems enhance usability and ethical considerations. Students will be encouraged to discuss how xAI could improve MT workflows and end-user trust.	Slides 8 to 15

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 4: Ethical Challenges in MT	The teacher will lead a discussion on the ethical challenges associated with MT systems, emphasising fairness, bias, and accountability. Key discussion points include: - Bias in training data and its impact on MT outputs (e.g., gender bias, cultural insensitivity). - Fairness in evaluating MT systems for diverse languages and user groups. - The ethical responsibility of researchers and developers in minimising harm caused by MT errors. Students will review real-world examples of ethical challenges in MT and brainstorm strategies to address them. The teacher will emphasise the importance of ethical evaluation practices to ensure MT systems are designed and deployed responsibly.	Slides 16 to 24

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 5: User-Centric Evaluation	The teacher will introduce the concept of user-centric evaluation in MT systems, focusing on the importance of designing evaluations that reflect real-world use cases. Key points to discuss: - Understanding the needs and expectations of different user groups. - Designing evaluations that prioritise usability and relevance to the target audience. - Balancing technical metrics with user feedback to gain holistic insights into MT quality. Students will be encouraged to share examples of how MT systems have been evaluated in user-centric contexts. The teacher will guide a discussion on the challenges and benefits of incorporating user perspectives into MT evaluations.	Slides 25 to 35

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 6: Designing Comprehensive Evaluation Frameworks	The teacher will outline the key components of a comprehensive evaluation framework for MT systems, emphasising the importance of aligning evaluation methods with system goals. The session will cover: - Combining human evaluation, automatic metrics, and user feedback. - Selecting appropriate evaluation tools and platforms for different scenarios. - Creating detailed guidelines and workflows to ensure consistency and reliability. Students will work in groups to design a preliminary evaluation framework based on a hypothetical MT system. The teacher will provide feedback on their designs and facilitate a discussion on potential improvements.	Slides 36 to 50

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 7: Evaluating MT for Low-Resource Languages	The teacher will discuss the unique challenges of evaluating MT systems for low-resource languages. Key discussion points include: - The lack of sufficient training data and its impact on MT quality. - The importance of culturally and linguistically sensitive evaluation methods. - Strategies for designing evaluations when resources (e.g., evaluators or reference texts) are limited. Students will analyse case studies of MT systems for low-resource languages and brainstorm ways to improve evaluation practices in these contexts. The teacher will emphasize the ethical and practical importance of addressing the needs of low-resource languages in MT research.	Slides 51 to 60

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 8: Reflective Discussion and Recap	The teacher will facilitate a reflective discussion to summarise the key concepts covered in the lesson. Discussion prompts include: - What are the most significant challenges in creating transparent and ethical MT evaluations? - How can xAI contribute to more reliable and user-centric evaluations? - What have you learned about balancing technical and user-centric evaluation methods? Students will be asked to share their insights and propose actionable steps for applying the lesson's concepts to their own MT evaluation projects. The teacher will conclude by highlighting the importance of collaboration and continuous learning in advancing ethical and transparent MT practices.	Slides 61 to 70

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Closing Remarks	The teacher will conclude the session by summarizing the key takeaways: - The critical role of transparency and ethical considerations in MT evaluation. - How user-centric evaluation enhances the relevance and impact of MT systems. - Strategies for addressing challenges in low-resource language evaluations. Students will be encouraged to continue exploring these topics and reflect on how they can apply the concepts in their academic or professional work. The teacher will provide a brief overview of the next session, emphasizing its focus on advanced evaluation techniques and practical applications.	-

F.8 Lesson Plan 8 - Lab

Lesson Plan 8 - Lab

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 1: Warming Up	The teacher will begin the lesson by asking students how they are doing. The teacher will organise students into groups for the next activity.	-

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 2: Ethics - Case Studies	Each group will receive a case study, which is a report of an MT system containing several performance-based claims. **Activity Steps:** 1. Students will be given 10 minutes to analyse the case studies. They should read through the case studies and identify potential flaws in quality assessment, methodology, and any questionable claims. 2. Students should consider ways to accurately report findings or suggest improvements to avoid the pitfalls described in the case studies. 3. After reading the case studies, identifying flaws, and making corrections, each group will present the original text, highlighting all the issues they found. The teacher will project their own annotated version of the case studies' issues, enabling a comparison with students' findings to discuss where they may have agreed or differed.	Case Study Documents

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 3: TBL - Performance Case Study	The teacher will conclude the previous activity and conduct a case study with the students as a group. **Activity Steps:** 1. The teacher will read aloud a case study that focuses on Performance. 2. The teacher will prompt students to identify aspects of the Triple Bottom Line (TBL), specifically relating to Performance, Planet, and People. 3. The teacher will note down the students' suggestions and then compare these suggestions with an annotated version, highlighting how the paper may or may not have addressed the TBL. 4. Finally, the teacher will summarise the discussion to prepare students for the next task, where they will be divided into groups to analyse different papers on the TBL for Topic 4.	Case Study Documents

ACTIVITY	METHODOLOGICAL PROCEDURES	TOOLS
Topic 4: TBL - Case Studies	The teacher will divide students into groups. **Activity Steps:** 1. One group will read a summary of Henderson et al. (2020), while another group will read a summary of Shterionov and Vanmassenhove (2021). 2. Students will read their assigned summary and, based on the TBL, identify contributions related to "People," "Planet," and "Performance." 3. Students are then expected to brainstorm ideas on how to adapt these contributions to MT quality assessment. Students are encouraged to look up the full version of the paper online to verify further details if needed. After 15 minutes, students will present their paper, describe the identified TBL aspects, and suggest ways to integrate the paper's contributions into MT quality assessment.	Paper Summaries, Online Research
Topic 5: Closing Remarks	The teacher will conclude the session by asking students to complete the end questionnaire and provide feedback on the session.	Attachment I - End Questionnaire

Appendix G

Phase 5 - Pretest Questionnaire

Research Introduction

This study is part of a research project titled "Teaching Translation Quality Assessment Hybrid and Online: From Pedagogical Rationale to Analysis, Design, Implementation, and Evaluation," funded by the School of Applied Language and Intercultural Studies (SALIS) at Dublin City University (DCU). This research is led by João Lucas Cavalheiro Camargo (joo.cavalheirocamargo2@mail.dcu.ie) under the supervision of Dr. Joss Moorkens (DCU) and Dr. Sheila Castilho (DCU).

This study focuses on implementing and training a syllabus for Machine Translation Quality Assessment (MTQA) that has been developed and refined based on feedback from lecturers in Natural Language Processing and Translation Technology. The goal is to evaluate and improve the syllabus through a training program in which participants engage in discussions, activities, and assessments.

As a participant, you will be involved in a training program delivered by the researcher, who will also act as the instructor. You will participate in various activities and discussions, and your interactions will provide valuable feedback for improving the training.

Ethical Aspects

- No anticipated risks from involvement in this research study; anonymity and confidentiality are ensured.

- Each participant is assigned a number before data processing to ensure anonymity.
- Data will be securely disposed of five years after collection.
- Participation is voluntary, and you may withdraw at any time without repercussions.
- If concerns arise, the Faculty Research Ethics Committee (F-REC) at DCU can be contacted at `hss.frec@dcu.ie`.

Informed Consent Please confirm your agreement with the following statements by ticking the corresponding box:

- I have read and understood the Plain Language Statement.
- I am aware of the observation in educational settings.
- I am aware that my answers are confidential within legal limitations.
- I understand that participation is voluntary and can be withdrawn at any time.
- I accept that I will not receive financial compensation.

[ ] **I consent to participate in this research.**

1. **What is your email?** _____
2. **What degree are you currently pursuing?**

   Bachelor's Degree

   Master's Degree

   PhD

   Other (Please specify): _____

3. **What is your field of study?**

   Natural Language Processing (NLP)

Translation Technology

Other (Please specify): _____

4. What are your expectations from this training?

---

5. Are you familiar with evaluation metrics?

No

Yes

6. Which evaluation metrics are you familiar with?

Automatic Evaluation Metrics

Human Evaluation Metrics

I am not familiar with either

7. If familiar with automatic metrics, list them:

---

8. If familiar with human metrics, list them:

---

9. How confident are you in using automatic evaluation metrics?

1 - Not Confident at All: I have very little or no understanding of automatic evaluation metrics and do not feel confident in using them.

2 - Slightly Confident: I have some basic knowledge of automatic evaluation metrics but feel unsure about applying them effectively.

3 - Moderately Confident: I am reasonably confident in my understanding of automatic evaluation metrics and can use them with some level of assurance, although I may encounter difficulties.

4 - Very Confident: I am quite confident in both my knowledge of and ability to use automatic evaluation metrics effectively.

5 - Extremely Confident: I am very knowledgeable and feel highly confident in my ability to effectively use automatic evaluation metrics in various contexts.

10. How confident are you in using human evaluation metrics?

1 - Not Confident at All: I have very little or no understanding of human evaluation metrics and do not feel confident in using them.

2 - Slightly Confident: I have some basic knowledge of human evaluation metrics but feel unsure about applying them effectively.

3 - Moderately Confident: I am reasonably confident in my understanding of human evaluation metrics and can use them with some level of assurance, although I may encounter difficulties.

4 - Very Confident: I am quite confident in both my knowledge of and ability to use human evaluation metrics effectively.

5 - Extremely Confident: I am very knowledgeable and feel highly confident in my ability to effectively use human evaluation metrics in various contexts.

11. Explain the use of automatic evaluation metrics in your own words:

12. Explain the use of human evaluation metrics in your own words:

Appendix H

Phase 5 - Posttest Questionnaire

1. What is your email? _____

2. What technical aspects do you consider essential when developing Machine Translation evaluation, and why?

_____

_____

3. What ethical aspects do you consider essential when developing Machine Translation evaluation, and why?

_____

_____

4. What do you think would be the consequences if machine translation assessments were lenient or omitted altogether?

_____

_____

5. How confident are you in using automatic evaluation metrics?

1 - Not Confident at All: I have very little or no understanding of automatic evaluation metrics and do not feel confident in using them.

2 - Slightly Confident: I have some basic knowledge of automatic evaluation metrics but feel unsure about applying them effectively.

3 - Moderately Confident: I am reasonably confident in my understanding of automatic evaluation metrics and can use them with some level of assurance, although I may encounter difficulties.

4 - Very Confident: I am quite confident in both my knowledge of and ability to use automatic evaluation metrics effectively.

5 - Extremely Confident: I am very knowledgeable and feel highly confident in my ability to effectively use automatic evaluation metrics in various contexts.

6. How confident are you in using human evaluation metrics?

1 - Not Confident at All: I have very little or no understanding of human evaluation metrics and do not feel confident in using them.

2 - Slightly Confident: I have some basic knowledge of human evaluation metrics but feel unsure about applying them effectively.

3 - Moderately Confident: I am reasonably confident in my understanding of human evaluation metrics and can use them with some level of assurance, although I may encounter difficulties.

4 - Very Confident: I am quite confident in both my knowledge of and ability to use human evaluation metrics effectively.

5 - Extremely Confident: I am very knowledgeable and feel highly confident in my ability to effectively use human evaluation metrics in various contexts.

7. How confident are you in planning MTQA evaluations?

1 - Not Confident at All: I have very little or no understanding of MTQA evaluation planning and do not feel confident in creating or structuring evaluation plans.

2 - Slightly Confident: I have a basic understanding of MTQA elements but feel unsure about integrating them effectively into a structured evaluation plan.

3 - Moderately Confident: I am reasonably confident in my knowledge of MTQA evaluation planning and can create plans with some assurance, though I may still encounter challenges.

4 - Very Confident: I am quite confident in my knowledge of MTQA evaluation planning and feel capable of developing structured evaluation plans effectively.

5 - Extremely Confident: I am very knowledgeable and feel highly confident in my ability to design and implement comprehensive MTQA evaluation plans across various contexts.

8. Did this training meet your expectations?

---

9. Was there anything missing from this training that you expected to learn or experience?

---

10. Is there anything else you would like to share about your experience with this training?

Appendix I

Inferential Analysis: Bayesian Linear Mixed Modelling

This section contains the full results of a Bayesian LLM. It contains the fitness of the model, results brokwn down by groups and all variables. Models were estimated separately for each variable, HEConfidence and AEMConfidence. A first-order autoregressive covariance structure [AR(1)] provided the best balance of fit and parsiomony when compared to alternative structures such as compound symmetry or unstructured covariance. Fixed effects included Time (Pretest, Mid-test, Posttest), Background (NLP vs TS), and their interaction. The results presented below offer a more robust evaluation of the effectiveness of the training by testing whether the increase in confidence observed descriptively was statistically reliable across the sample.

I.0.0.1 Confidence in Human Evaluation

The linear mixed model analysis revealed a significant main effect of Time on HEConfidence, $\mathrm{F}(2, 25.67) = 17.73$, $p < .001$., indicating substantial changes in participants' confidence levels across the three measurement points. In contrast, the main effect of Background ($\mathrm{F}(1, 14.05) = 1.75$, $p = .207$) and the Time $\times$ Background interaction ($\mathrm{F}(2, 25.67) = 0.36$, $p = .701$) were not statistically significant, suggesting that confidence increased similarly regardless of participants' disciplinary

Effect	Num df	Den df	F	Sig. (p)
Intercept	1	14.051	516.74	$< .001$
Time	2	25.665	17.73	$< .001$
Background	1	14.051	1.75	.207
Time × Background	2	25.665	0.36	.701

Table I.1: Summary of fixed effects from the linear mixed model for HEConfidence. The table reports each predictor, the numerator and denominator degrees of freedom used in the F-test, the F-value, and the associated significance level (p-value).

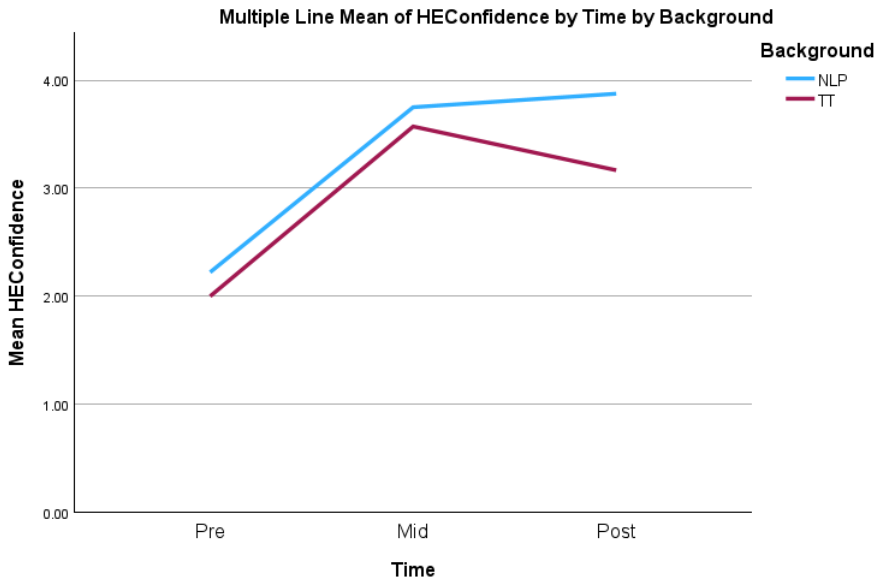


Figure I.1: Mean *HEConfidence* across time, by participant background (NLP vs TT).

background.

Estimates of marginal means (Table I.2) showed a significant increase in HEConfidence from Pre (M = 2.11, SE = 0.202) to Mid (M = 3.65, SE = 0.227), with a slight decline at Post (M = 3.53, SE = 0.237), although still markedly higher than at baseline. Pairwise contrasts revealed that the Pre–Mid increase was statistically significant (p=.011p=.011), while the Mid–Post change was not.

Model diagnostics supported the robustness of the fitted model. The first-order autoregressive (AR[1]) structure was retained based on model comparison, with an AIC of 126.51 and a BIC of 131.73, providing a favourable balance between fit and parsimony. The variance estimates for the AR(1) diagonal ($\sigma^2 = 0.684$) and random intercepts ($\tau^2 = 0.088$) were small and stable, further indicating good model

convergence and suitability.

When broken down by background (Table I.3), participants from both NLP and TT groups followed a similar trend, showing large improvements from Pre to Mid and then levelling off or slightly declining by Post. The NLP group showed marginally higher overall scores, but the non-significant interaction confirms that the pattern of change over time was consistent between the two groups, as it can be observed in figure I.2.

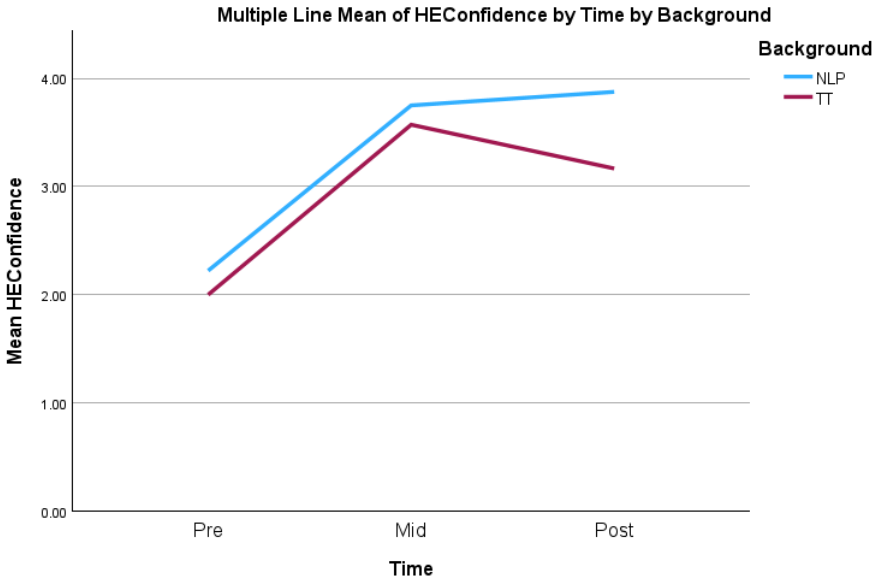


Figure I.2: Mean *HEConfidence* across time, by participant background (NLP vs TT).

As a next stage, calculating the estimated marginal means provides model-adjusted average confidence scores across timepoints. These means offer a clearer picture of how HEConfidence changed throughout the study, accounting for individual variation and the fixed effects included in the model. Table I.2 presents the estimated means for each timepoint, while Table I.3 displays the same estimates disaggregated by participant background.

The estimated marginal means in Table I.2 show a marked increase in HEConfidence from the Pre-test (M = 2.11, SE = 0.20) to the Mid-test (M = 3.65, SE = 0.23), indicating a substantial gain in confidence during the first half of the training. At Post-test, the mean remained high (M = 3.53, SE = 0.24), suggesting that this

Time	Mean	SE	95% CI Lower	Upper
Pre	2.111	0.202	1.704	2.519
Mid	3.653	0.227	3.194	4.112
Post	3.526	0.237	3.048	4.004

Table I.2: Estimated marginal means of HEConfidence by Time

increase was maintained through to the end of the course. The relatively narrow confidence intervals across all timepoints further support the stability of these estimates. While there was a slight decrease from Mid to Post, the difference was minimal and within the margin of error.

Time	Background	Mean	SE	95% CI Lower	Upper
Pre	1 (NLP)	2.000	0.278	1.439	2.561
	2 (TT)	2.222	0.293	1.631	2.814
Mid	1 (NLP)	3.557	0.332	2.887	4.227
	2 (TT)	3.750	0.311	3.122	4.377
Post	1 (NLP)	3.193	0.358	2.471	3.915
	2 (TT)	3.859	0.310	3.232	4.485

Table I.3: Estimated marginal means of HEConfidence by Time and Background

The estimated marginal means disaggregated by background (Table I.3) indicate broadly similar trajectories in HEConfidence for both groups. At Pre test, participants from the TT group reported slightly higher confidence (M = 2.22, SE = 0.29) than those from the NLP group (M = 2.00, SE = 0.28), though both groups began at similarly low levels. By Mid test, both groups showed substantial improvement, with TT participants reaching a mean of 3.75 (SE = 0.31) and NLP participants at 3.56 (SE = 0.33). At Post-test, the TT group maintained a slightly higher mean (3.86, SE = 0.31) compared to the NLP group (3.19, SE = 0.36).

Despite these numerical differences, the lack of a statistically significant Time × Background interaction suggests that the pattern of change over time was not meaningfully different between the groups. Both cohorts demonstrated a strong early increase in confidence followed by a period of stabilisation or slight regression.

Effect	Num df	Den df	F	Sig. (p)
Intercept	1	13.874	549.31	$< .001$
Time	2	26.962	9.52	$< .001$
Background	1	13.874	5.66	.032
Time × Background	2	26.962	0.36	.700

Table I.4: Linear mixed model results for AEMConfidence

I.0.0.2 Confidence in Automatic Evaluation Metrics

The linear mixed model for the variable AEMConfidence revealed a significant main effect of Time, F(2, 26.96) = 9.52, p ¡ .001, indicating that confidence levels changed significantly across the three timepoints. A significant main effect of Background was also observed, F(1, 13.87) = 5.66, p = .032, suggesting that participants' disciplinary background influenced their overall confidence in using AEMs. However, the Time × Background interaction was not significant, F(2, 26.96) = 0.36, p = .700, indicating that while the groups differed in overall confidence, their trajectories of change over time followed a similar pattern.

Time	Mean	SE	95% CI Lower	Upper
Pre	2.378	0.201	1.972	2.784
Mid	3.391	0.226	2.935	3.848
Post	3.437	0.234	2.965	3.910

Table I.5: Estimated marginal means of AEMConfidence by Time

The estimated marginal means presented in Table I.5 show a steady increase in AEMConfidence across the three timepoints. Confidence rose from Pre (M = 2.38, SE = 0.20) to Mid (M = 3.39, SE = 0.23), and continued to increase slightly at Post (M = 3.44, SE = 0.23). The relatively narrow confidence intervals suggest these estimates are stable and reliable. Unlike the more abrupt pattern observed for HEConfidence, the growth in AEMConfidence appears more gradual and sustained across the duration of the training.

The estimated marginal means broken down by background (Table I.6) show a consistent difference between the NLP and TT groups across all timepoints. At Pre-test, the NLP group reported higher confidence in using AEMs (M = 2.56, SE

Time	Background	Mean	SE	95% CI Lower	Upper
Pre	NLP (1)	2.556	0.291	1.967	3.144
	TT (2)	2.200	0.277	1.641	2.759
Mid	NLP (1)	3.760	0.309	3.136	4.384
	TT (2)	3.023	0.330	2.356	3.690
Post	NLP (1)	3.825	0.308	3.204	4.447
	TT (2)	3.049	0.352	2.338	3.761

Table I.6: Estimated marginal means of AEMConfidence by Time and Background

= 0.29) compared to the TT group (M = 2.20, SE = 0.28), likely due to their computational background. This gap persisted throughout the training, with the NLP group maintaining higher scores at both Mid (M = 3.76 vs. 3.02) and Post (M = 3.83 vs. 3.05).

Despite these between-group differences, the lack of a significant Time × Background interaction indicates that the pattern of change over time was similar across both groups. That is, both NLP and TT participants demonstrated gains in AEMConfidence over time, but the initial disparity in confidence levels remained stable rather than widening or narrowing. This suggests that the TT group improved at a comparable rate to the NLP group, though from a lower starting point.

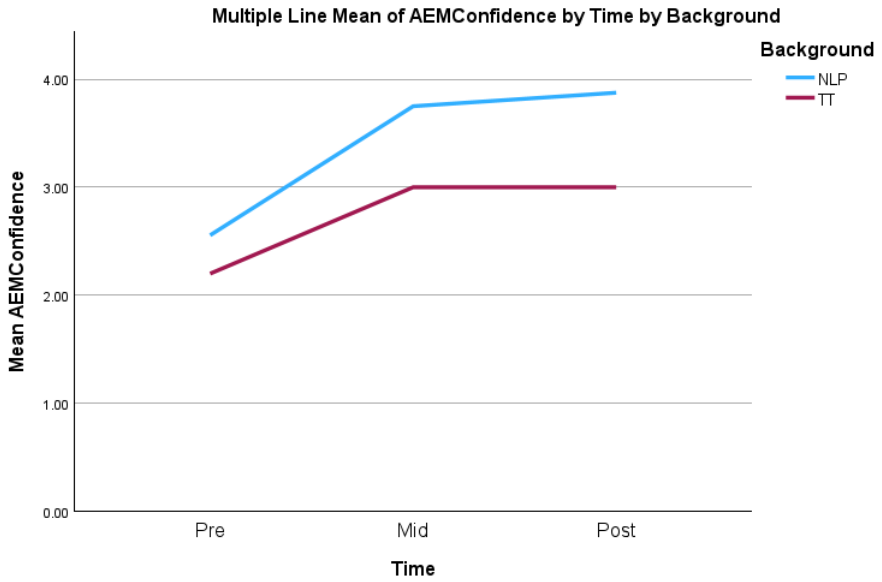


Figure I.3: Mean AEMConfidence across time, by participant background (NLP vs TT).

To summarise, figure I.3 shows a consistent upward trend for both the NLP and

TT groups, with no evidence of divergence or convergence between them. While the NLP group reported higher confidence at all three timepoints, the nearly parallel slopes reinforce the non-significant interaction effect observed in the model. This suggests that although initial familiarity with AEMs differed by background, both groups benefited similarly from the training in terms of confidence growth.

I.0.0.3 Summary of results

Across both models, participants demonstrated statistically significant increases in confidence over time in their use of HE and AEMs. Gains in HEConfidence were particularly pronounced between the Pre and Mid timepoints, followed by a levelling off at Post, while AEMConfidence increased more gradually but consistently across all three stages. The presence of significant Time effects in both models confirms that the training was effective in improving participants' self-reported confidence levels.

Although participants' disciplinary background (NLP vs TT) influenced their overall confidence levels—particularly in the domain of AEMs—the trajectory of change over time did not significantly differ between groups. This suggests that while initial familiarity varied, both groups engaged productively with the training and achieved comparable gains in confidence.

Appendix J

Questionnaire responses

Participant Code	Group	Question 4 - What are your expectations or desired outcomes from participating in the training?	Question 11 - Explain the use of AEMs with your own words.	Question 12 - Explain the use of HE metrics with your own words.
P4S1	TT	Familiarise myself with what evaluating MT implies and how it is performed		
P4S2	NLP	To be able to understand how MT evaluation works and develop an automatic multimodal AEM	The automatic evaluation metrics are computer programs that deals with identify how well the MT output is correlated with human translation	Its evaluation performaned by humans based on predefined criteria.

Participant Code	Group	Question 4 - What are your expectations or desired outcomes from participating in the training?	Question 11 - Explain the use of AEMs with your own words.	Question 12 - Explain the use of HE metrics with your own words.
P4S3	TT	Learn more about evaluation of MT, how it is carried out, and things to notice		
P4S4	TT	Get to know more about the HE involved in MTQA and the prafctical evaluation methods that can be applied in real scenarios	Counting the similarity of n-grams between the reference translation and MT output and then give a score	
P4S5	NLP	Learning more about how MT works (basics or broadly) and in particular about the human and automatic metrics that can be used to see how well the MT system is doing. I also would like to see from a linguistic view, how appropriate MT systems are and ethical concerns.	To see how well the system is doing. Different metrics see how well different aspects are doing.	

Participant Code	Group	Question 4 - What are your expectations or desired outcomes from participating in the training?	Question 11 - Explain the use of AEMs with your own words.	Question 12 - Explain the use of HE metrics with your own words.
P4S6	TT	Learning more about Automated and Human Evaluation of Machine Translation	I know it gives a score to the MT according to its similarity with a reference translation made by a human translator	
P4S7	NLP	MT		
P4S8	TT	To be able to evaluate MT systems and do inferences from the reports	They are used for evaluation and assessment of the machine translations	
P4S9	TT	I hope to know more about MT evaluation processes	I recently know that chrf is the metric that compares the machine translation output with the reference character by character to evaluate the similarity between the two, and generate a score of the MT quality	

Participant Code	Group	Question 4 - What are your expectations or desired outcomes from participating in the training?	Question 11 - Explain the use of AEMs with your own words.	Question 12 - Explain the use of HE metrics with your own words.
P4S10	NLP	Understand the theories and techniques, then learn how to build up a machine translation model, fine tune and evaluate it.	Do the calculation, use existed data and results to compare the results from different models.	
P4S11	TT	Having deep knowledge of how to use AEMs to evaluate the machine translation	It is used to measure the similarity between machine translation and the human one by n-grams	
P4S12	NLP	Looking to improve the skill of evaluations and reporting it in a better way	To see how good the model is performing for the desired task	
P4S13	NLP	General understanding of machine translation and improving llm (fine-tuning) for low resource languages	To reduce time too evaluate and reduce human error and bias	High quality but time consuming process. But with some human bias and errors.

Participant Code	Group	Question 4 - What are your expectations or desired outcomes from participating in the training?	Question 11 - Explain the use of AEMs with your own words.	Question 12 - Explain the use of HE metrics with your own words.
P4S14	N/A	I would like to know more about this issue		Try to understand about people and societies.
P4S15	TT	Improvement in my understanding of MT evaluation. Awareness of the ethical issues of MT evaluation. Furthering my knowledge in the Translation Technology sphere. Ability to help evaluation with Irish language MT in this emerging sector.	Using a pre trained or untrained system to evaluate how close an MT output compare to a gold standard reference translation. Characters and sentence length are used in different metrics. TER analyses the maount of post editing needed on an output compared to a reference translations	Using human translators to evaluate a translation output against a MT output
P4S16	NLP	I am hoping to learn more about machine translation in any form		

Participant Code	Group	Question 4 - What are your expectations or desired outcomes from participating in the training?	Question 11 - Explain the use of AEMs with your own words.	Question 12 - Explain the use of HE metrics with your own words.
P4S17	NLP	To know something extra and different and figured what I like eventually		
P4S18	NLP	I expect to gain an intermediate understanding of machine translation assessment and evaluation		
P4S19	TT	Developing and increasing my knowledge about Quality Assessment		

Appendix K

Thematic Analysis Occurrences per Unit

Subtheme	Occurrence	Percentage
Teaching Delivery	1	1.28%
Difficulty levels	2	2.56%
Contextual Factors	0	0.00%
Transferability to other NLP tasks	1	1.28%
Purpose of MT use	11	14.10%
Caution with MT use	7	8.97%
Reflections on MT use	14	17.95%
Human-Centred & Critical Perspectives	15	19.23%
UX/UI	0	0.00%
Metrics	1	1.28%
Concept of Quality	1	1.28%
Evaluation Frameworks & Criteria	6	7.69%
Ethical Considerations	0	0.00%
Standardised norms	1	1.28%
Platforms	1	1.28%
Transferability	0	0.00%
Pedagogical Aspects	0	0.00%
Expectations over Content	17	21.79%

Table K.1: Occurrences and percentages of subthemes coded in Unit 1.

Unit 1

Subtheme	Occurrence	Percentage
Teaching Delivery	2	2.99%
Difficulty levels	4	5.97%
Contextual Factors	1	1.49%
Transferability to other NLP tasks	1	1.49%
Purpose of MT use	1	1.49%
Caution with MT use	0	0.00%
Reflections on MT use	4	5.97%
Human-Centred & Critical Perspectives	7	10.45%
UX/UI	11	16.42%
Metrics	7	10.45%
Concept of Quality	0	0.00%
Evaluation Frameworks & Criteria	19	28.36%
Ethical Considerations	0	0.00%
Standardised norms	0	0.00%
Platforms	8	11.94%
Transferability	1	1.49%
Pedagogical Aspects	1	1.49%
Expectations over Content	0	0.00%

Table K.2: Occurrences and percentages of subthemes coded in Unit 2.

Unit 2

Unit 3

Unit 4

Subtheme	**Occurrence**	**Percentage**
Teaching Delivery	0	0.00%
Difficulty levels	0	0.00%
Contextual Factors	0	0.00%
Transferability to other NLP tasks	0	0.00%
Purpose of MT use	0	0.00%
Caution with MT use	0	0.00%
Reflections on MT use	0	0.00%
Human-Centred & Critical Perspectives	2	5.41%
UX/UI	0	0.00%
Metrics	0	0.00%
Concept of Quality	0	0.00%
Evaluation Frameworks & Criteria	35	94.59%
Ethical Considerations	0	0.00%
Standardised norms	0	0.00%
Platforms	0	0.00%
Transferability	0	0.00%
Pedagogical Aspects	0	0.00%
Expectations over Content	0	0.00%

Table K.3: Occurrences and percentages of subthemes coded in Unit 3.

Subtheme	**Occurrence**	**Percentage**
Teaching Delivery	10	12.82%
Difficulty levels	2	2.56%
Contextual Factors	2	2.56%
Transferability to other NLP tasks	2	2.56%
Purpose of MT use	0	0.00%
Caution with MT use	11	14.10%
Reflections on MT use	18	23.08%
Human-Centred & Critical Perspectives	0	0.00%
UX/UI	0	0.00%
Metrics	1	1.28%
Concept of Quality	0	0.00%
Evaluation Frameworks & Criteria	0	0.00%
Ethical Considerations	15	19.23%
Standardised norms	0	0.00%
Platforms	0	0.00%
Transferability	2	2.56%
Pedagogical Aspects	5	6.41%
Expectations over Content	10	12.82%

Table K.4: Occurrences and percentages of subthemes coded in Unit 4.

Appendix L

Faculty Research Ethics Committee Approval



Faculty of Humanities & Social Sciences
DUBLIN CITY UNIVERSITY
28 February 2023

CONFIRMATION OF RESEARCH ETHICS APPROVAL FOR A PROJECT

Application Reference: **DCU-FHSS-2023-015**

Project Title: **Teaching Translation Quality Assessment Hybrid and Online: From Pedagogical Rationale to Analysis, Design, Implementation and Evaluation**

Project contact(s): **joo.cavalheirocamargo2@mail.dcu.ie**

Let this letter certify that the proposed project identified above has been reviewed by the *Humanities & Social Sciences Faculty Research Ethics Committee* (F-REC) and has been approved as a low-risk project. The application was found to comply with university requirements and best practices for research ethics, and with GDPR guidelines and requirements where personal data is processed in the project.

A copy of the application, including appended documents related to participant consent, is archived under the reference above. Queries about this project's approval may be directed to the F-REC Chair.

Sincerely,

Dr Dónal Mulligan
donal.mulligan@dcu.ie

Chair, Faculty Research Ethics Committee
Faculty of Humanities & Social Sciences
Dublin City University

Dámh na nDaonnachtaí agus na nEolaíochtaí Sóisialta
Ollscoil Chathair Bhaile Átha Cliath